"Responding to Asian Diversity"

DH2022

DIGITAL HUMANITIES 2022

Conference Abstracts

July 25-29, 2022
Tokyo, Japan

Online
The University of Tokyo

Australasian Association for Digital Humanities (aaDH)
Association for Computers and the Humanities (ACH)
Canadian Society for Digital Humanities / Société canadienne des humanités numériques (CSDH/SCHN)
centerNet
Digital Humanities Association of Southern Africa (DHASA)
Digital Humanities im deutschsprachigen Raum (DHd)
European Association for Digital Humanities (EADH)
Humanistica, L'association francophone des humanités numériques/digitales (Humanistica)
Japanese Association for Digital Humanities (JADH)
Red de Humanidades Digitales (RedHD)
Taiwanese Association for Digital Humanities (TADH)

Digital Humanities 2022

Conference Abstracts

The University of Tokyo, Japan

25-29 July 2022

Primary Encoder: Yifan Wang
Encoders and Proofreaders:
Tomohiro Murase
Kiyonori Nagasaki
Yoshihiro Sato
Shintaro Seki
Illustration: hata design
Available online at: https://dh2022.adho.org/

Published by DH2022 Local Organizing Committee

DIGITAL HUMANITIES 2022
RESPONDING TO ASIAN DIVERSITY

25-29 July 2022, Toshi Center Hotel, Tokyo, Japan and Fully Online (Zoom)

The international DH conference has played an eminent role not only in providing a forum for the presentation of completed research results, but also in creating a forum for the exchange of knowledge and research collaboration in the nascent stage of research, and it was therefore judged to be as desirable as possible to hold the meeting face-to-face, even if it involves some difficulties.

However, due to the continuous emergence of new variants, COVID-19 shows no signs of ending, and countries around the world are faced with intermittent and severe restrictions in various aspects, including medical care, education, research, and freedom of travel and movement. In the midst of this fluid situation, it is impossible to choose the hybrid meeting as it was envisioned.

In light of this situation, the LOs decided to change the DH2022 to be held entirely online, giving priority to providing equal and safe opportunities for all those who wish to participate. By providing a safe, respectful, and principled environment online for participants from around the world, regardless of the differences in their circumstances, we hope to contribute to the realization of ADHO's principles.

The details of the online meeting plan will be expanded from the online meeting portion of the current plan to the full plan. We will do our best to accommodate ADHO members from all over the world in the program and other aspects of the conference.

International Program Committee

Co-chairs
Ikki Ohmukai 大向一輝
Taizo Yamada 山田太造

Tully Barnett
Jen-Jou Hung 洪振洲
Ian Milligan
Emmanuel Ngué Um
Thomas Padilla
Miriam Peña Pimentel
Petr Plecháč
Nadezhda Povroznik
Mmasibidi Setaka
Georg Vogeler

Local Organizing Committee

Chair
Masahiro Shimoda 下田正弘

Satoko Fujiwara 藤原聖子
Kaori Karasawa 唐沢かおり
Mari Kobayashi 小林真理
Masato Kobayashi 小林正人
Mareshi Saito 齋藤希史
Kyoko Sengoku-Haga 芳賀京子
Akira Takagishi 高岸輝

Secretariat

Head
Kiyonori Nagasaki 永崎研宣

Michihiko Aono 青野道彦
Daigo Isshiki 一色大悟
Takahiro Kato 加藤隆宏
Koichi Takahashi 高橋晃一
Toru Tomabechi 苫米地等流

Conference Volunteers

Yuta Hashimoto 橋本雄太
Akihiro Kameda 亀田尭宙
Naoki Kokaze 小風尚樹
Tomohiro Murase 村瀬友洋
Satoru Nakamura 中村覚
Jun Ogawa 小川潤
Yoshihiro Sato 左藤仁宏
Shintaro Seki 関慎太朗
Ayako Shibutani 渋谷綾子
Yifan Wang 王一凡
Yoichiro Watanabe 渡邉要一郎
Fumi Yao 八尾史

SNS team

Setsuko Yokoyama 横山説子
Yuan Li 李媛
Yifan Wang 王一凡

CFP Translation

Christoph Beutelspacher - Deutsch
Lidia Ponce de la Vega - Español
Paolo Saporito - Italiano
Marie Schvarzman - Français

Session chairs

Mari Agata 安形麻理
Elisa Beshero-Bondar
Fabio Ciotti
Gregory Crane
James Cummings
Thomas Dabbs
Maciej Eder
Emi Hamana 浜名恵美
Yuta Hashimoto 橋本雄太
Bor Hodošček
Jen-Jou Hung 洪振洲
Yuri Ishida 石田友梨
Yoichi Iwasaki 岩崎陽一
Diane Jakacki
Akihiro Kameda 亀田尭宙
Nadia Kanawaga 金川ナディア
Akihiro Kawase 河瀬彰宏
Asanobu Kitamoto 北本朝展
Naoki Kokaze 小風尚樹
Ryuta Komaki 小牧龍太
Wakako Kumakura 熊倉和歌子
Glen Layne-Worthey
Chao-Lin Liu 劉昭麟
Luis Meneses
So Miyagawa 宮川創
Hajime Murai 村井源
Natsuko Nakagawa 中川奈津子
Satoru Nakamura 中村覚
Chifumi Nishioka 西岡千文
Terhi Nurmikko-Fuller
Emily Ohman
Elena Pierazzo
Nadezhda Povroznik
Christof Schöch
Martina Scholger
Walter Scholger
Susan Schreibman
Michael Sinatra
Peter Stokes
Tomoji Tabata 田畑智司
Koichi Takahashi 高橋晃一
Menno Van Zaanen
Christian Wittern
Taizo Yamada 山田太造
Hilo Yamamoto 山元啓史
Mishio Yamanaka 山中美潮
Keiji Yano 矢野桂司
Natsuko Yoshiga 吉賀夏子

Conference Sponsors

COMMERCIAL SPONSORS

Bronze sponsor: Gale
Bungaku-Report
Flx style co.,LTD.
Infomage
Google
Metainfo
Stanford University Press

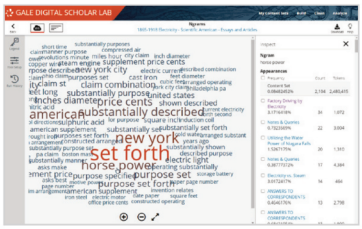
FUNDINGS

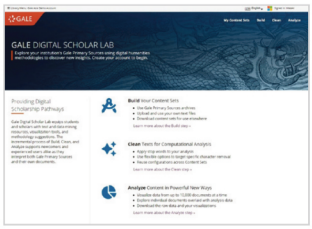
JSPS KAKENHI Grant-in-Aid for Scientific Research (A) "Inheritance and development of the digital knowledge base of Buddhist studies" JP19H00516 (PI: Professor Masahiro Shimoda, The University of Tokyo)
JSPS Program for Constructing Data Infrastructure for the Humanities and Social Sciences (Historiographical Institute, The University of Tokyo) (PI: Keiko Hongo)

INSTITUTIONAL SPONSORS

BUKKYO DENDO KYOKAI
International Insitute for Digital Humanities, Tokyo
The Mitsubishi Foundation
National Insitute of Japanese Literature, Tokyo

EXPERIENCE THE NEW
GALE DIGITAL SCHOLAR LAB



With rich content sets, a cloud-based platform, intuitive analysis tools, and a built-in learning center, ***Gale Digital Scholar Lab*** represents the evolution of our continued support of digital humanities

GALE

www.gale.com/digital-scholar-lab

Google

.supDigital

Keynote

The journey to make billions of Japanese historical documents accessible

Dr. Tarin Clanuwat ทาริน คลานุวัฒน์
Google Brain



Japan is a unique country with one of the world's most distinctive cultures. Over a thousand years of Japanese culture and knowledge are preserved inside historical documents, currently inaccessible to 99.99% of Japanese people to whom the documents are unreadable. They are written in Kuzushiji, or cursive Japanese, which looks like patterns on paper for most people. Hence, in order to preserve the culture of Japan, collaboration and participation of the general public is indispensable. In this keynote, I will describe what Kuzushiji is, what makes it so important and how machine learning can help us unlock and shed some light on our rich history. I will also talk about the efforts of many organizations in tackling this problem and how we collaborate despite completely different approaches to the same goal: to make Japanese historical documents accessible to everyone.

Profile

Dr. Tarin Clanuwat is a senior research scientist at Google Brain based in Tokyo. She received her PhD in classical Japanese literature from the Graduate school of Letters Arts and Sciences at Waseda University, where she specialized in Kamakura-era Tale of Genji commentaries. In 2018, while she was a project assistant professor at ROIS-DS Center for Open Data in the Humanities (CODH), she developed an AI-based kuzushiji recognition model called KuroNet. In 2019, she hosted a Kaggle machine learning competition for kuzushiji recognition which attracted over 300 machine learning researchers and engineers from around the world. In order to make Japanese historical documents more accessible to the general public, Tarin also developed “miwo” AI Kuzushiji recognition smartphone application.

Her kuzushiji recognition research and applications won several awards including Yamashita Memorial Research Award from the Information Processing Society of Japan, research paper award from Japan Society for Digital Archive Conference and Excellent award from Vitalizing Local Economy Organization by Open Data & Big Data.

Keynote

Harnessing Chinese Historical Writings for Digital Humanities

Dr. Jieh Hsiang 項潔
National Taiwan University



About nine hundred years ago Zheng Qiao (鄭樵, 1104~1162 CE), a Chinese scholar of the Southern Song Dynasty, ventured on an ambitious project of "collecting all books into one book". The backdrop of his motivation was that China had a uniform written language for more than a millennium at the time (more than two at present) with a long tradition of keeping historical records even with the dynastic changes. To illustrate his idea, Zheng Qiao wrote a book called Tongzhi (通志, Comprehensive Treatises) in which he condensed pre-Tang Chinese history into 200 volumes (about 5,000,000 characters). For example, in the Outline of Anomalies (《通志。災祥略》), he listed all the anomalies such as solar eclipses and earthquakes that he could find from historical records and used their frequencies to argue that, contrary to common belief, such natural phenomena did not correlate to how benign the ruler was (and therefore were not warnings from heaven). Zheng Qiao was way ahead of his time and Tongzhi had largely been ignored by later scholars.

The emergence of digital humanities has cast new lights on Zheng Qiao's vision.

In this talk, we shall give a brief overview of the issues and challenges facing the digitization of Chinese historical writings, together with the current state of digitally available searchable full-text. We shall describe the importance of tagging (of time, person, place, and objects), how they are done, and how they can be used to connect documents to create textual contexts. We shall present a theory of (textual) context discovery and show how it can be used to build retrieval systems aimed at discovering, analyzing, and visualizing contexts hidden among documents. A platform, DocuSky, which enables humanities scholars to build their own personal context discovery retrieval systems without the help of IT specialists will also be presented. We shall complete the presentation by giving a short discussion of how Zheng Qiao might have done were he living in the DH era.

Profile

Jieh Hsiang is a Distinguished Professor in Computer Science of the National Taiwan University. He is also the director of the NTU Research Center for Digital Humanities, the first such center in the Sinophone world. In his pre-DH life, he received a Ph.D. in computer science from the University of Illinois at Urbana-Champaign and worked mainly in automated theorem proving, in particular term rewriting systems. After returning to Taiwan in 1993, Jieh Hsiang started to work with with historians and anthropologists and initialized the digitization of Taiwanese cultural heritage at his university. Through the years, he and his team built over 30 large scale digital libraries of Chinese/Taiwanese historical archives, all of which utilize a context-discovery retrieval methodology that is designed for scholarly use of digital archives. His team also developed DocuSky, a personal DH platform for humanities scholars to process, annotate, analyze and visualize data and build their own searchable text databases without the help of an IT specialist. The NTU Digital Library of Buddhist Studies, a comprehensive

bibliographic on-line database that he has been in charge of since 2005, attracts more than 11,000 users around the world each day.

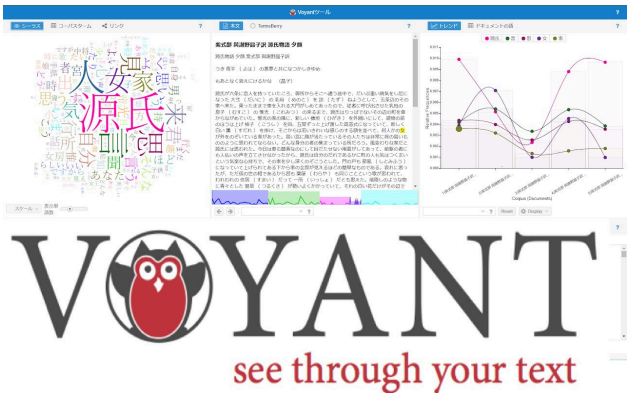
Jieh Hsiang was the author/editor of six books on digital humanities in Chinese. Being the first of its kind, the books made significant impact in promoting DH in the Sinophone world. He organized the first Conference on Digital Archives and Digital Humanities in 2009, which becomes the first annual DH conference in East Asia. He also helped establishing the Taiwanese Association for Digital Humanities in 2016, and the first Chinese language journal on DH, the Journal of Digital Archives and Digital Humanities, which was inaugurated in 2018. Jieh Hsiang served 6 years as the University Librarian of NTU and 8 years as the Director of NTU Press. Before returning to Taiwan, he was a full professor in Computer Science at Stony Brook University.

Keynote

The Winner of the Triennial Antonio Zampolli Prize: Voyant Tools



Dr. Geoffrey Rockwell
University of Alberta



Voyant Tools is an in-browser suite of tools for the study of electronic texts. You upload a text to the Voyant server or use one already indexed and then you have access to coordinated set of analysis and visualization tools. Voyant will show you trends over the text or allow you to compare different texts in a corpus. It will provide textual statistics and word lists that you can export for use in other tools.

Voyant was designed by humanists for the interpretation of texts. It doesn't give you answers, it allows you to explore small or large text collections intuitively. As it is free and runs in the browser on our server it is used for introducing students to text analysis. To learn more about the computer-assisted interpretation of texts with Voyant see our book:

Rockwell G. & S. Sinclair. Hermeneutica. MIT Press, 2016

Profile

Dr. Geoffrey Rockwell is a Professor of Philosophy and Digital Humanities at the University of Alberta, Canada. He is also the Director of the Kule Institute for Advanced Study and Associate Director of AI for Society signature area. He received his MA and PhD in Philosophy from the University of Toronto where he also worked in computing as a Senior Instructional Technology Specialist. From 1994 to 2008 he was at McMaster University where he directed the Humanities Media and Computing Centre and helped found the Department of Communication Studies and Multimedia.

Rockwell publishes on video games, textual visualization, text analysis, ethics of technology and on digital humanities including a co-edited book on Right Research: Modelling Sustainable Research Practices in the Anthropocene (Open Book Publishers, 2021) and a co-authored book Hermeneutica: Computer-Assisted Interpretation in the Humanities (MIT Press, 2016). He is co-developer of Voyant Tools (voyant-tools.org), an award winning suite of text analysis tools. He is a co-organizer of the Replaying Japan conference, co-editor of the Journal of Replaying Japan, and President of the Canadian Society for Digital Humanities.

Pre-Conference Workshops and Tutorials

Visual Analysis of Printed Illustrations using Computer Vision
Bergel, Giles Edward; Dutta, Abhishek 27
Navigating and Processing Data from the TEI with XSLT
Beshero-Bondar, Elisa Eileen; Scholger, Martina; Nagasaki, Kiyonori 28
Equity, Diversity and Inclusion for Digital Humanists
Bordalejo, Barbara; O'Donnell, Daniel; Woods, Nathan 30
Introduction to DraCor – Programmable Corpora for Digital Drama Analysis
Boerner, Ingo; Fischer, Frank; Milling, Carsten; Trilcke, Peer; Sluyter-Gäthje, Henny 32
Writing a multilayered article for the Journal of Digital History
Clavert, Frédéric; Elisar, Ori; Pfeiffer, Mirjam; Guerard, Elisabeth 35
Tutorial on Fuzzy String Matching with DeezyMatch
Coll Ardanuy, Mariona; Hosseini, Kasra; Nanni, Federico; Vitale, Valeria 37
Biographical Data in a Digital World 2022 (BD 2022) Workshop
Daza, Angel; Fokkens, Antske; Hadden, Richard; Hyvönen, Eero; Koho, Mikko; Wandl-Vogt, Eveline 39
Workshop: HathiTrust Research Center's Extracted Features 2.0 Dataset
Dubnicek, Ryan; Christie, Jennifer; Kudeki, Deren; Layne-Worthey, Glen; Walsh, John A.; Downie, J. Stephen 43
A picture is worth a thousand words: Image analysis for the Digital Humanities
James, Stuart; Aubry, Mathieu; Van Noord, Nanne; Garcia, Noa; Impett, Leonardo 45
Data and Algorithms in Critical Aging Studies
Karadkar, Unmil; Kriebernegg, Ulla; Sawchuk, Kim; Taipale, Sakari; Ivan, Loredana 48
Literary Text Analysis with Spyral Notebooks, a Notebook Environment Companion to Voyant Tools
Land, Kaylin Catherine; Rockwell, Geoffrey; MacDonald, Andrew; Tchoc, Bennett Kuwan; Damasah, Elliot 50
Making 3D-scans more mobile-friendly and increasing online audience reach: Introduction to manual retopology
Leelasorn, Angel 51
Getting Started with the Advanced Research Consortium
Liebe, Lauren; Mandell, Laura; Tarpley, Bryan 52
Scholarly writing and editing with the text editor Stylo
Mellet, Margot Lise; Fauchié, Antoine; Vitali-Rosati, Marcello 54
Ugarit: Translation Alignment Technologies for Under-resourced Languages
Palladino, Chiara; Yousef, Tariq; Shamsian, Farnoosh; Kanagawa, Nadia 55
From Concepts to Textual Phenomena and Back: Operationalization in the Digital Humanities
Pichler, Axel; Krautter, Benjamin; Pagel, Janis; Andresen, Melanie 57
Hands-on Introduction to eScriptorium, an Open-Source Platform for HTR
Stokes, Peter Anthony; Stökl Ben Ezra, Daniel 58
Text and data mining for East Asian sources in classical Chinese
Sturgeon, Donald 61
Git for Humanists: Versioning Research and Code
Tagliaferri, Lisa 63
How to Set Up a Web Server for Teaching and Research in the Humanities
Tagliaferri, Lisa 63

Panels

Computer Vision for the Study of Printers' Ornaments and Illustrations in European Hand-Press Books
Bahier-Porte, Christelle; Bergel, Giles; Dutta, Abhishek; Fournel, Thierry; Thomas, Drew; Vial-Bonacci, Fabienne; Wilkinson, Hazel; Zisserman, Andrew; Denis, Loïc; Emonet, Rémi; Habrard, Amaury; Ventresque, Vincent; Gautrais, Thomas 65
The European Open Science Cloud (EOSC) and Its Implications for the Digital Humanities and Social Sciences
Barbot, Laure; Gray, Edward; Fischer, Frank; Broeder, Daan; Ďurčo, Matej; Kleemola, Mari; Thiel, Carsten; König, Alexander 68

Modelling and Operationalizing Concepts in Computational Literary Studies
Brandes, Phillip; Dennerlein, Katrin; Jacke, Janina; Marshall, Sophie; Pielström, Steffen; Schneider, Felix *70*
The Politics of Digital Humanities Infrastructure and Sustainability
Burkert, Mattie; Moore, Shawn; Gil, Alex; Liebe, Lauren; Nicosia, Marissa; Otis, Jessica; Wikle, Olivia; Williamson, Evan Peter; Becker, Devin *73*
Temporal Topologies: Inflecting the telling and the told of historical narratives
Drucker, Johanna; Dörk, Marian; Morini, Francesca; LaCelle-Peterson, Nathaniel; Rinderlin, Jonas; Barker, Elton; Rosol, Christoph; Wintergruen, Dirk *76*
Global Perspectives on Critical Infrastructure and the Digital Humanities Lab
Hannah, Matthew Nathan; Connell, Sarah; Dodd, Maya; Ope-Davies (Opeibi), Tunde; Povroznik, Nadezhda; Rittenhouse, Brad *79*
The Ethical Considerations of Diverse DH Pedagogy
Licastro, Amanda Marie; Stringfield, Ravynn K.; Earhart, Amy; Losh, Elizabeth *82*
Digital debating cultures: Communicative practices on Reddit
Messerli, Thomas C.; Dayter, Daria; Bohmann, Axel; Donlan, Lisa; Maccori Kozma, Gustavo; Leuckert, Sven; Liimatta, Aatu; Mahler, Hanna; Massanari, Adrienne; McConnell, Kyla; Tosin, Rafaela *83*
SpokenWeb: Curating Literary Sound in a Digital Environment
O'Driscoll, Michael; Luyk, Sean; Kroon, Ariel; Morrison, Zachary; Ambarani, Tejas; Miya, Chelsea *87*
CLIP and beyond: Multimodal and Explainable Machine Learning in the Digital Humanities
Offert, Fabian; Impett, Leonardo; Al Moubayed, Noura; Cetinic, Eva; Bell, Peter; Smits, Thomas; Leone, Anna; Watson, Matthew; Winterbottom, Tom; Kluvanec, Dan; Lawrence, Dan; Kosti, Ronak; Wevers, Melvin; Lefranc, Lith *89*
Books' Impact in Digital Social Reading: Towards a Conceptual and Methodological Framework
Pianzola, Federico; Viviani, Marco; Fossati, Alessandro; Boot, Peter; Fialho, Olivia; Koolen, Marijn; Neugarten, Julia; Van Hage, Willem Robert; Rebora, Simone; Herrmann, J. Berenike; Messerli, Thomas C.; Jorschick, Annett; Sharma, Srishti *94*
Dynamics of Culture: Tracing Discourse using Computational Methods
Quinn, William Reed; Messina, Cara Marta; Connell, Sarah; Blankenship, Avery *98*
The (Im)Possibilities of Multilingual DH in Theory and Practice: Translation, Metadata, Pedagogy
Raynor, Cecily; Ponce de la Vega, Lidia; Guénette, Marie-France; Kim, Eric; Brata Roy, Samya; Dombrowski, Quinn *101*
Cultural Landscapes in Emerging Digital Scholarship: The Search of Conceptual and Computational Frameworks
Streiter, Oliver; Chuang, Tyng-Ruey; Zhan, Hanna Yaqing; Hara, Shoichiro; Hung, Ying-Fa; Jang, Jr-Jie; Lee, Cheng-Jen; Mu, Yu-Chia Monica; Wang, Chia-Hsun Ally; Wang, Yu-Huang *104*

Long Presentations

A Computational Approach to Epistemology in Poetry of the Long Eighteenth Century
Algee-Hewitt, Mark Andrew *109*
Gender Assignment as an Event: a contemporary approach to adequately depict historical gender categories
Andrews, Tara Lee; Ebel, Carla; Deierl, Marin *111*
Online Readership and Perceptions of Genres Over Time
Antoniak, Maria; Walsh, Melanie; Mimno, David *113*
Manifesting the manifesto: DH and the climate crisis
Baillot, Anne; Gil Fuentes, Alexander; Glover, Kaiama L; Peaker, Alicia; Roeder, Torsten; Scholger, Walter; Walton, Jo Lindsay *115*
DFG 3D-Viewer – Development of an infrastructure for digital 3D reconstructions
Bajena, Igor Piotr; Dworak, Daniel; Kuroczyński, Piotr; Smolarski, René; Münster, Sander *117*
Representing uncertainty and cultural bias with Semantic Web technologies
Baroncini, Sofia; Daquino, Marilena; Pasqual, Valentina; Tomasi, Francesca; Vitali, Fabio *120*
Beyond the Tracks: connecting people, places and stations to re-assess the impact of rail in Victorian Britain
Beelen, Kaspar; McDonough, Katherine; Lawrence, Jon; Rhodes, Josh; Wilson, Daniel C.S. *122*

Machine Learning May Be the Future, but Can It Be the Past? What Machine Learning Systems May Mean for the Historical Concept of Provenance
Benito-Santos, Alejandro; Doran, Michelle; Edmond, Jennifer; Therón, Roberto *124*
Processing tangles in the *Frankenstein Variorum*
Beshero-Bondar, Elisa Eileen; Borgia, Mia; Chan, Jacqueline; Viglianti, Raffaele *126*
Who's In and Who's Out: 10 Years On
Bleeker, Elli; Beelen, Kaspar; Chambers, Sally; Koolen, Marijn; Melga-Estrada, Liliana; Van Zundert, Joris J. *128*
Exploring Lexical Diversities
Blombach, Andreas; Evert, Stephanie; Jannidis, Fotis; Pielström, Steffen; Konle, Leonard; Proisl, Thomas *130*
Support and Relationship Patterns in Endometriosis Narratives
Bologna, Federica; Thalken, Rosamond; Mimno, David; Wilkens, Matthew *134*
Distant reading of Russian Soviet diaries (Prozhito Database)
Bonch-Osmolovskaya, Anastasia; Vorobieva, Viktoria; Kriukov, Artem; Podriadchikova, Maria *136*
Linguistic Injustice in Multilingual Technologies: arTenTen and esTenTen as case studies
Bordonaba-Plou, David; Jreis-Navarro, Laila M. *140*
TikTok Cover Dances as Folkloric Practice: Pose Estimation and the Study of Variation in K-Pop Choreography across Short-Form Social Media Videos
Broadwell, Peter; Tangherlini, Timothy R. *141*
Tools as Epistemologies in DH? A Corpus-Based Exploration
Burghardt, Manuel; Luhmann, Jan; Niekler, Andreas *144*
Evaluation of Multilingual BERT in a Diachronic, Multilingual, and Multi-Genre Corpus of Bibles
Calvo Tello, José; De la Rosa, Javier *147*
Archive of the Digital Present (ADP), COVID-19 Period: Collecting and Visualizing Metadata of Online Literary Events Hosted in Canada, March 2020 - September 2021
Camlot, Jason; Neugebauer, Tomasz; Berrizbeitia, Francisco; Joseph, Ben; Bustamante, Alexandre; Gandham, Sukesh *151*
Textual, Metrical and Musical Stylometry of the Trouvères Songs
Camps, Jean-Baptiste; Chaillou, Christelle; Mariotti, Viola; Saviotti, Federico *155*
Data Diversity in handwritten text recognition: challenge or opportunity?
Camps, Jean-Baptiste; Pinche, Ariane; Stutzmann, Dominique; Vernet, Marguerite; Vidal-Gorène, Chahan *160*
Distant reading of handwritten mid-nineteenth century Ottoman population registers
Can, Yekta Said; Kabadayi, M. Erdem *165*
Named Entity Disambiguation for the Qing Shilu without Manually Labeled Data
Chao, Jo-Yu; Huang, Shi-Yun; Hsieh, Hsin-Yi; Tsai, Richard Tzong-Han *168*
Digital Narrative of Buddhist Cave Temples: A Case Study of Niche 28 of Huangze Temple in Guangyuan, Sichuan
Chen, Wu Wei; Zuo, Lala *171*
A Novel Semi-supervised Framework to Identify Military Documents: A Quantitative Analysis on Military Records in *Ming Shi-Lu*
Chen, You-Jun; Hsieh, Hsin-Yi; Tsai, Richard Tzong-Han *172*
Linked Open Dictionaries (2015-2022): Achievements, Experiences and Challenges with respect to LOD Technology in Linguistics and the Philologies
Chiarcos, Christian; Ionov, Maxim; Fäth, Christian *176*
Building ETCSANS: The Electronic Text Corpus of Syntactically Annotated Neo-Sumerian
Chiarcos, Christian; Page-Perron, Emilie *179*
Computational approaches to literary periodization: an experiment in Italian narrative of 19th and 20th century
Ciotti, Fabio *181*
Building a journal for the digital era: the Journal of Digital History
Clavert, Frédéric; Fickers, Andreas *183*
New ways to read Greek and Persian epic and to explore diverse cultures
Crane, Gregory; Shamsian, Farnoosh; Babeu, Alison; Tauber, James; Wegner, Jacob *185*
Good Grief! Encoding, Quantifying, and Analyzing *Peanuts* in the Classroom
Croxall, Brian *188*

Multilinguals Write Back: Modeling Language, Politics and Identity in Philippine Social Media
Cruz, Frances Antoinette; Kestemont, Mike *189*
Strategies for DH awareness-raising activities: the challenges of DH Awards
Cummings, James *191*
Emotion courses in German historical comedies and tragedies
Dennerlein, Katrin; Schmidt, Thomas; Wolff, Christian *193*
Database Design and Identity: A Compromised Infrastructure
Earhart, Amy *197*
One word to rule them all: understanding word embeddings for authorship attribution
Eder, Maciej; Šeļa, Artjoms *199*
Measuring Keyness
Evert, Stephanie *202*
Historical Research meets Semantic Interoperability: The Documentation System SYNTHESIS and its Application in Art History Research
Fafalios, Pavlos *205*
Using Digital Tools to Detect Cross-Language Allusions in Voltaire
Gawley, James *207*
Reforming the 'Eng Lit' canon: Measuring the myths and realities of English literary studies in India through a computational analysis of university curricula.
Ghosh, Arjun *209*
An Adaptive Methodology: Machine Learning and Literary Adaptation
Glass, Grant *210*
Optical Character Recognition for Complex Scripts: A Case-study in Cuneiform
Gordin, Shai; Romach, Avital *212*
The many faces of theory in DH: Toward a dictionary of theoreticians mentioned in DH
Gutiérrez de la Torre, Silvia E.; Burghardt, Manuel; Niekler, Andreas; Kleymann, Rabea *215*
Out of the Slaughterhouse: The Birth of the Modern Detective Story Corpus
Hammond, Adam; Stern, Simon *218*
Voices Speaking To and About One Another: Introducing the Project Dialogism Novel Corpus
Hammond, Adam; Vishnubhotla, Krishnapriya; Mohammad, Saif M.; Hirst, Graeme *220*
Realizing a multilingual tool for best legal and ethical practices in DH research: The ELDAH Consent Form Wizard as a model for community-driven internationalization
Hannesschläger, Vanessa; Kuzman-Slogar, Koraljka; Scholger, Walter *224*
CCVG Data: A Unique, Curated, and Searchable Chinese Village Dataset for Chinese Study Scholars
He, Daqing; Ma, Rongqian; Zheng, Ruoyun; Zhang, Haihui *225*
Structuring the Management of Research Data - Reflections on Requirements and Service Concepts in Research Data Management in the Humanities
Helling, Patrick *227*
Making Research Data FAIR. Seriously? - Reflections on Research Data Management in the Computational Literary Studies
Helling, Patrick; Jung, Kerstin; Pielström, Steffen *230*
Building an OCR Pipeline for a Republican Chinese Entertainment Newspaper
Henke, Konstantin; Arnold, Matthias *233*
Townsend & Sons, Account Book Manufacturer's Business Guide and Works Manual: 19th Century Primary Manuscript Source and TEI Encoding
Hermsen, Lisa; Walker, Rebekah *236*
Interactive Visualization of Indigenous Territory: Mapping of the Attikamegues in Historical Maps
Herold, Nastasia; Blicher Christensen, Mathilde; Jänicke, Stefan *238*
Using Word Embeddings for Validation and Enhancement of Spatial Entity Lists
Herrmann, J. Berenike; Byszuk, Joanna; Grisot, Giulia *239*
How big can a static site be? Staticizing a census database
Holmes, Martin; Newton, Greg *241*
PRISMS: a new platform for digital Book History
Huber, Alexander; Huber, Emma *243*

Everyday memory: A computational analysis of changing relation between past and present in Dutch newspapers in the twentieth Century
Huijnen, Pim *246*
Mining for clean energy: a machine learning approach to historicized sentiment mining of fossil fuel discourse in the Netherlands
Huijnen, Pim; Plets, Gertjan; Verheul, Jaap *249*
Participatory Action Research for a Digital Humanities research project: Investigating Open GLAM in the context of Social Movement Archives
Humbel, Marco *252*
Linked Data Approach for Studying Parliamentary Speeches and Networks of Politicians in Finland 1907-2021
Hyvönen, Eero; Leskinen, Petri; Sinikallio, Laura; Drobac, Senka; Tuominen, Jouni; Elo, Kimmo; La Mela, Matti; Koho, Mikko; Ikkala, Esko; Tamper, Minna; Leal, Rafael; Kesäniemi, Joonas *254*
On the Road to Freedom: Network models of interviews of Czechoslovak respondents in the optics of audio-emotional analysis and computational linguistics.
Iashchenko, Anatoly Vladimirovich *256*
Modelling Gender Diversity – Research Data Representation Beyond the Binary
Illmer, Viktor J.; Poggel, Lisa; Diehr, Franziska; Drury, Lindsey *257*
Abstractness/ Concreteness as Stylistic Features for Authorship Attribution
Ivanov, Lubomir *262*
Studying Signs of Use in Medieval Manuscripts: data collection through annotations
Johnson, David F.; Girard, Paul; Simard, Benoît *264*
Standards, the Standards-Making Process, and their Relevance to Stylometry
Juola, Patrick *267*
Quantitative Analysis of Changes in Race and Role of Characters in Hollywood Movies over Time
Kawase, Akihiro; Isogai, Kana *268*
Quantitative Analysis of Gendered Assumptions in a Nineteenth-Century Women's Encyclopedia
Ketzan, Erik; Hagen, Thora; Jannidis, Fotis; Witt, Andreas *270*
On Digitizing Historic Music Storage Media For Computational Analysis
Khulusi, Richard; Fricke, Heike; Fuhry, David; Piontkowitz, Vera; Focht, Josef *274*
Book Barcoding: A Framework for the Visual Collation and Woodblock Tracking of Japanese Printed Books
Kitamoto, Asanobu *276*
Emotions and Literary Periods
Konle, Leonard; Kröncke, Merten; Jannidis, Fotis; Winko, Simone *278*
Accuracy is not all you need
Kristensen-McLachlan, Ross Deans; Lassen, Ida Marie S.; Enevoldsen, Kenneth; Hansen, Lasse; Nielbo, Kristoffer L. *281*
Semantically-Grounded Generative Modeling of Chinese Landscapes
Kulyabin, Mikhail; Kosti, Ronak; Bell, Peter *284*
From Roland to Conan: First results on the corpus of French literary fictions (1050-1920)
Langlais, Pierre-Carl; Camps, Jean-Baptiste; Baumard, Nicolas; Morin, Olivier *285*
Newspaper Navigator: Reimagining Digitized Newspapers with Machine Learning
Lee, Benjamin Charles Germain *289*
Between Interactive Fictions and Visual Novels: Diversity of Agency in Videoludic Novels
Lescouet, Emmanuelle; Dumoulin, Pierre Gabriel *291*
Gender and Cultural Diversity in Chinese Children's Picture Books: A Data-led Analysis of Bestselling Modern Titles
Li, Yi; Terras, Melissa; Li, Yongning *292*
DH Pedagogies in the Global Souths
Licastro, Amanda Marie; Roy, Dibyadyuti; Esprit, Schuyler Kirshten *295*
Rethinking the Advanced Research Consortium: Disciplinary Restructuring and Linked Open Data
Liebe, Lauren; Mandell, Laura *296*
Discovering Civil Disputes Hidden in the Court Judgment Documents for Applications in Social Studies and Legal Informatics
Liu, Chao-Lin; Liu, Yi-Fan; Liu, Wei-Zhi; Lin, Hong-Ren *298*

Digital Resource Aggregation: Giving New Life to Multi-source Cultural Data
Liu, Rui; McKay, Dana; Buchanan, George *300*
Mining the Native American Authored Works in HathiTrust for Insights
Lu, Kun; Heaton, Raina; Orr, Raymond; Vetter, Alyssa; Dubnicek, Ryan; Magni, Isabella *301*
Applying LERA for collating witnesses of The Tale of Kiều, a Vietnamese poem written in Nôm script
Luu, Thi Kim Hanh; Pöckelmann, Marcus; Ritter, Jörg; Molitor, Paul *302*
A 3-D analytic framework of humanistic objects: data modeling paradigms, computational analysis and close reading
Maeir, Noam; Keydar, Renana *305*
Towards Adaptive Retrieval Systems for Intertextuality Research: a case study on Biblical Intertextuality
Manjavacas Arevalo, Enrique *307*
Aplicación de las Humanidades Digitales a los estudios sobre Teatro de la Antigüedad Clásica
Martínez Nieto, Roxana Beatriz *308*
From Modern to Medieval: Detecting and Visualizing Entities in Manuscripts of Marco Polo's Devisement du Monde
Meinecke, Christofer; Wrisley, David Joseph; Jänicke, Stefan *310*
A Five-Star Model for Linked Humanities Data Usability
Middle, Sarah *313*
Exploring a cultural “filter bubble” in artwork databases of two large museums of fine arts
Minster, Sara; Kizhner, Inna; Zhitomirsky-Geffet, Maayan *315*
EthicsBot: Provoking Ethical Reflection on AI
Mousavi, Emad; Verdini, Paolo; Wang, Jingwei; Barnard, Sara; Rockwell, Geoffrey *317*
Digital Resistance to Asian-American Hate during COVID-19: Study of Photography and Art on Instagram
Nanditha, Narayanamoorthy *318*
#METOO IN INDIA: MISOGYNY AND THE EMERGENCE OF THE MEN’S RIGHTS MOVEMENT ON TWITTER
Nanditha, Narayanamoorthy *320*
Are Digital Humanities platforms sufficiently facilitating diversity in research? *A study of Transkribus free processing requests.*
Nockels, Joseph Hiliary; Terras, Melissa; Gooding, Paul; Muehlberger, Guenter; Stauder, Andy *321*
Small Data projects/Big Data research: contemporary problems and historical solutions
O'Donnell, Daniel; Woods, Nathan; Bordalejo, Barbara *323*
Co-reference networks for dramatic texts: Network analysis of German dramas based on co-referential information
Pagel, Janis *326*
Knowledge organization of the Hong Kong Martial Arts Living Archive to capture and preserve intangible cultural heritage
Picca, Davide; Adamou, Alessandro; Hou, Yumeng; Egloff, Mattia; Kenderdine, Sarah *329*
The case for DH in Literary scholarship
Pierazzo, Elena *332*
Minor Labels: Detecting Genre in Pitchfork Reviews, a "Metamodularity" Network Analysis
Porter, J.D.; Varner, Stewart *333*
Digitizing Derrida’s Concept of Dissemination: From Returntocinder.com to Databyss.org
Reeder, Jake *335*
Locating a national collection through audience research
Rees, Gethin *337*
Telescopic reading: Synthesizing meaning from reading at different scales
Ringler, Hannah; Argamon, Shlomo *339*
Archiviz: A Tool for the Interactive, Visual Exploration of Digital Archives
Rittenhouse, Brad; Michney, Todd; Acosta, Ines *340*
Web Services for Voyant: LINCS, Voyant and NSSI
Rockwell, Geoffrey Martin; Hervieux, Natalie; Zafar, Huma; Land, Kaylin; MacDonald, Andrew; Barbosa, Denilson; Frizzera, Luciano; Ilovan, Mihaela; Brown, Susan *342*
From *Cyclopaedia* to *Encyclopédie*: Using Machine Translation and Sequence Alignment to Identify Encyclopedia Articles across Languages
Roe, Glenn; Olsen, Mark; Morrissey, Robert *344*
Establishing parameters for stylometric authorship attribution of 19th-century Arabic books and periodicals
Romanov, Maxim; Grallert, Till *346*

The Moderniſa Project: Orthographic Modernization of Spanish Golden Age Dramas with Language Models
De la Rosa, Javier; Cuéllar, Álvaro; Lehmann, Jörg *348*
Democratizing Poetry Corpora with Averell
De la Rosa, Javier; Díaz, Aitor; Pérez, Álvaro; Ros, Salvador; González-Blanco, Elena *351*
Developing the Japanese Visual Media Graph: An Open Knowledge Graph for Researchers Working on Japanese Anime, Manga and Otaku Culture
Roth, Martin; Pfeffer, Magnus; Kacsuk, Zoltan *354*
Digital Dating and its Discontents: AI, Masculinity and Consent
Roy, Dibyadyuti; Dahiya, Lavanya; Dahiya, Vasundhra *355*
The Interpretation of Dreams: A Case Study in Virtual Reality Filmmaking and the Remediation of Psychoanalytic Theory
Sack, Graham Alexander *357*
Comparing Symbolism Across Asian Cultural Contexts Using Graph Similarity Measures
Sartini, Bruno; Vogelmann, Valentin; Van Erp, Marieke; Gangemi, Aldo *358*
Inner- and Intra-genre Citation Patterns in Film
Schneider, Stefanie *361*
Annotating 3D Scholarly Editions
Schreibman, Susan; Papadopoulos, Costas *364*
Measuring Space in German Novels
Schumacher, Mareike Katharina *365*
The model of choice.Using pure CRF- and BERT-based classifiers for gender annotation in German fantasy fiction
Schumacher, Mareike Katharina; Flüh, Marie; Lemke, Marc *368*
Systems of Sentencing in Medieval Inquisitorial Records: semantic text modelling as a platform for computational analysis
Shaw, Robert L. J. *370*
Poetry as Error. A 'Tool Misuse' Experiment on the Processing of German Language Poetry
Sluyter-Gäthje, Henny; Trilcke, Peer *372*
Digital Humanities and Replication. Ingredients for a Love Story – Experiences from the '(Re)counting the Uncounted' Project
Stapel, Rombert *375*
Describing Handwriting in Context
Stokes, Peter Anthony *378*
Towards a crowdsourced linked open knowledge base of East Asian historical sources
Sturgeon, Donald *380*
Intertextuality in Large-Scale East Asian and Western European Corpora
Tharsen, Jeffrey; Gladstone, Clovis *382*
Connecting Digital Systems for whom and by whom? Taking Stock of the Digital Research Infrastructures in China
Tsui, Lik Hang; Chen, Jing *385*
Object constitution in digital collections: An Ethnomethodological View
Türkoglu, Enes; Mertgens, Andreas *386*
Using Temporal Information on Topic Mining
Uno, Takeaki; Kobayashi, Ryota; Takedomi, Yuka; Hashimoto, Takako *388*
Reducing Redundancy Bias in Digital Library Collections
VandenBosch, Adrienne; Organisciak, Peter; Matusiak, Krystyna K *389*
Teaching Digital Scholarly Editing North and South Through Minimal Computing
Viglianti, Raffaele; Del Rio Riande, Gimena *391*
DeXTER: A post-authentic approach to heritage visualisation
Viola, Lorella *393*
Textbooks and space – a tale of two dimensions. A GIS analysis of cultural content of language textbooks.
Wacławik, Paulina *396*
Mining an Interpolated Commentary for Linguistic Markup
Waxman, Joshua *397*
The Many Voices of Du Ying: Revisiting the Disputed Writings of Lu Xun and Zhou Zuoren
Xie, Xin; Wang, Haining; Riddell, Allen *400*
Japanese Old Maps Online for Promoting Digital Humanities
Yano, Keiji; Natsume, Muneyuki; Imamura, Satoshi; Kamata, Ryo *404*

Measuring the Use of Tools and Software in the Digital Humanities: A Machine-Learning Approach for Extracting Software Mentions from Scholarly Articles
Zarei, Alireza; Seung-Bin, Yim; Fischer, Frank; Ďurčo, Matej; Wieder, Philipp 406
Transfer Learning for Olfactory Object Detection
Zinnen, Mathias; Madhu, Prathmesh; Bell, Peter; Maier, Andreas; Christlein, Vincent 409

Short Presentations

Analysis of the Gutenberg 42-line Bible types aided by type-image recognition
Agata, Mari; Agata, Teru 415
RDF-star-based Digital Edition of Travel Journals
Alassi, Sepideh; Rosenthaler, Lukas 416
On Epistemic Comparability and Challenges on Data Reuse:The Experience of READ-IT
Antonini, Alessio 417
Salience in Literary Texts: A Combined Approach to the Relevance of Passages
Arnold, Frederik; Fiechter, Benjamin; Gius, Evelyn; Jäschke, Robert; Martus, Steffen; Vauth, Michael 418
Sound Predicts Meaning: Sound Iconic Relations between Vowels' Formants and Emotional Tone in German and Japanese
Auracher, Jan; Menninghaus, Winfried; Scharinger, Mathias 421
The Seven Steps: Building the DiGA Thesaurus
Autiero, Serena; Elwert, Frederik; Moscatelli, Cristiano; Pons, Jessie 422
Global and local citation networks as a new paradigm for multiple viewpoint investigation in historical literature: a case study of the Rabbinic literature corpus
Ben-Gigi, Nati; Katzoff, Binyamin; Schler, Jonathan; Zhitomirsky-Geffet, Maayan 424
A new gesture-based browsing experience for art historical research
Bernasconi, Valentine 425
Diachronic Lexicon Induction via Literary Translations
Birkenes, Magnus Breder; Johnsen, Lars G.; Kåsen, Andre 427
Fifty Shades of Twilight: Computationally Comparing Collocations in Twilight and 50 Shades of Grey
Bordalejo, Barbara; Van Zundert, Joris J.; Neugarten, Julia 428
Keeping Kanji Alive: Lessons in Digital Sustainability
Bosse, Arno; Hibino Lory, Harumi 430
Revealing 'Invisible' Poetry by W. H. Auden through Computer Vision
Brenner, Simon; Frühwirth, Timo; Mayer, Sandra 431
Relating the Unread: A Data-Rich Approach to the Literary Canon and the "Great Unread"
Brottrager, Judith 432
VisColl 2.0 and VCEditor. A new model and tool in the quiver of codicologists and bibliographers
Campagnolo, Alberto; Porter, Dot; Emery, Doug; Perkins, Patrick; Ransom, Lynn 433
Nursing the Subaltern: Using Digital Prosopography to Explore the Transnational Makings of Filipino Nurses Since 1898
Capucao. Jr., Reynaldo Caasi 435
Take a sip of TEI and relax: a proposition for an end-to-end workflow to enrich and publish data created with automatic text recognition
Chagué, Alix; Scheithauer, Hugo; Terriel, Lucas; Chiffoleau, Floriane; Tadjo Takianpi, Yves; Romary, Laurent 437
The Great Transformation of the Clan System in Early China: A Social Network Analysis of Clan-sign Inscriptions from 1300 BC to 900 BC
Chen, Yuqi; Wang, Linxu 440
Machines Reading Maps: from text on maps to linked spatial data
Chiang, Yao-Yi; Holmes-Wong, Deborah; Kim, Jina; Li, Zekun; McDonough, Katherine; Simon, Rainer; Vitale, Valeria 444
The Postil Time Machine: The Lithuanian Lutheran Postils of the 16th Century
Chiarcos, Christian; Gelumbeckaite, Jolanta; Drach, Mortimer 445
Building infrastructure for annotating medieval, classical and pre-orthographic languages: the Pyrrha ecosystem
Clérice, Thibault; Jolivet, Vincent; Pilla, Julien 447

The Language of Reviewers: Sentiment, Ratings, and Style in Japanese-Language Amazon Video Reviews
Conroy, Melanie; Nishi, Hironori *449*
Scaling Up and Scaling Down: Expanding and Contracting in the Move to Linked Data
Crompton, Constance; Schwartz, Michelle; Dangoisse, Pascale; Lipski, Candice *451*
Una mirada digital a cuatro obras dramáticas de Federico García Lorca. Aportes de las técnicas de visualización a la interpretación de texto dramático
Dabrowska, Monika; Santa Maria, Teresa *452*
Word frequencies in authorship attribution: A simple tweak to improve performance
Eder, Maciej *453*
US-CHINA ZIGZAG: Semantic shifts and sentiment analysis of the US Congress speeches on US policy towards China and Taiwan
Elkobi, Jonathan *455*
What can stylometry and LIWC infer from Octavia E. Butler's use of 6+ letter words?
Essam, Bacem *457*
States of the text: fixation, legitimation and multimodal publication
Fauchié, Antoine *458*
An experimental attempt to use Transfer Learning for Named Entity Recognition in letters from the 19th and 20th century
Flüh, Marie; Lemke, Marc *459*
Language Model Pre-Training for Historical English: Approaches and Evaluation
Fonteyn, Lauren; Manjavacas Arevalo, Enrique *461*
Reconstruction of cultural memory through digital storytelling: a case study of Shanghai Memory project
Fu, Yaming; Mahony, Simon; Liu, Wei *462*
Building inclusivity into our digitization projects- a case study of digital collections in Mexico
Galina Russell, Isabel; Priani Saisó, Ernesto *464*
Having a Ball: A Linked Data Approach to Fancy Dress in Colonial Australia
Gatti, Tommy; Nurmikko-Fuller, Terhi; Pickering, Paul; Swift, Ben *466*
Crowdsourced Distributed Open Collaborative Courses (DOCC) for inclusive, self-regulated learning: A study on OERs in India
Ghosh, Sharanya; Das, Rajarshi *468*
Sentiment lexicons or BERT? A comparison of sentiment analysis approaches and their performance.
Grisot, Giulia; Rebora, Simone; Herrmann, Berenike *469*
Event annotation for literary corpora analysis
Grunspan, Claude; Mélanie, Frédérique; Barré, Jean; Chardon, Laurette; Galleron, Ioana; Naguib, Marco; Plancq, Clément; Seminck, Olga; Poibeau, Thierry *471*
Shared Tasks in Computational Literary Studies: Guideline Creation, Annotation and Text Generation for the Analysis of Narrative Levels
Guhr, Svenja Simone; Reiter, Nils; Zarrieß, Sina; Gius, Evelyn *472*
Towards a prosopographical ecosystem: modelling, design, and implementation issues
Hadden, Richard William James; Schlögl, Matthias; Vogeler, Georg *474*
Lexical Semantic Change in Literary Criticism
Haider, Thomas Nikolaus; Gittel, Benjamin *476*
“App-Solute News:” Comparison of Analog and Digital Mode in Newspaper Reading Between Intergenerational Teams
Hartinger, Teresa; Marinšek, Urša *478*
Crowdsourcing as Collaborative Learning: A Participatory Annotation Project for the Photographic Materials of Shibusawa Eiichi
Hashimoto, Yuta; Kim, Boyoung; Nakamura, Satoru; Kokaze, Naoki; Inoue, Sayaka; Shigehara, Toru; Nagasaki, Kiyonori *479*
An extensible Cor framework for textual publication and structure.
Hayward, Nicholas John *483*
Why arts and humanities publications get retracted: a topic modeling analysis of the retraction notices
Heibi, Ivan; Peroni, Silvio *484*
Patterns of Verb Usage in Immanuel Kant's Critical Writings
Heßbrüggen-Walter, Stefan; Fischer, Frank; Meier-Vieracker, Simon *486*

Worlding databases: A decolonising approach to the structuring and representation of data about global arts
Hidalgo Urbaneja, Maribel; Velios, Athanasios; Goodwin, Paul *488*
Modelling the relationship between morphosyntactic features and discourse relations in a multimodal corpus of primary school science diagrams
Hiippala, Tuomo; Haverinen, Jonas *489*
When context matters. How to explore a knowledge graph of heraldic communication and its contexts of use in medieval and early modern Europe with methods such as graph embedding
Hiltmann, Torsten; Schneider, Philipp *491*
Topic Modeling the Nineteenth-Century Poetry Canon
Houston, Natalie *492*
Bert-based Chinese Buddhist Cannon Citation Extraction Model Utilizing Prior Defined Regex Pattern and Data Augmentation
HUNG, JEN JOU; Wang, Yu-Chun *494*
Low-stakes activities for text analysis instruction in the undergraduate classroom
Isuster, Marcela Y. *495*
The Inevitability of a Reproducibility Crisis in the Digital Humanities
Isuster, Marcela Y.; Rod, Alisa B. *496*
The Linked Editorial Academic Framework: Creating an editorial environment for collaborative scholarship and publication
Jakacki, Diane Katherine; Brown, Susan; Cummings, James; Ilovan, Mihaela; Black, Carolyn *496*
Document similarity and topic clues. A historiographical study case
Jolivet, Vincent; Torres, Sergio *497*
Information Platform for Linked Data of Regional Historical Materials and its Agent Name Finding Process
Kameda, Akihiro; Goto, Makoto *499*
Processes and Practicalities in Developing and Sustaining a Text Mining Platform: Gale Digital Scholar Lab
Ketchley, Sarah; Ludwig, Jess *500*
After deplatforming: digital methods for documenting Twitter and YouTube moderation
De Keulenaar, Emillie Victoria; Kisjes, Ivan *501*
Victorian400: Colorizing Victorian Illustrations
Kim, Hoyeol *502*
An experiment in agent-based probabilistic city population reconstruction
Kisjes, Ivan; Van Wissen, Leon *503*
Data Diffraction: A Counternarrative to Integration in Digital Humanities Research
Kleymann, Rabea *505*
Contabilizar el comercio imperial: Analysis of early double-entry accounting books using the TEI/DEPCHA
Kokaze, Naoki; Fushimi, Takeshi; Nakamura, Yusuke *506*
Sound Iconicity and Digital Humanities. A Case Study of Spanish Golden Age Theatre
Kroll, Simon *508*
Toward an Affordances Approach to Literacy in the Digital Humanities
Kulkarni, Kavita *509*
Visualizing Academic Networks and Trends through Acknowledgements: Japanese Scholars in Islam-related Studies
Kumakura, Wakako; Sunaga, Emiko *510*
Diary of our initiatory journey on the continent of data citation in SSH
Larrousse, Nicolas; Gray, Edward; Concordia, Cesare *512*
A unified path from data to publications: Three French infrastructures in the COMMONS project
Larrousse, Nicolas; Pellen, Marie; Roux, Dominique; Baude, Olivier *513*
Electronic Literature: thinking a taxonomy
Lescouet, Emmanuelle; Vitali-Rosati, Marcello *514*
Digitizing and Recovering the Knowledge of Traditional Chinese Colour of the Nanjing Brocade
Lin, Wensi; Chen, Jing; Zhang, Mengyue; Wang, Jisheng; Li, Mengqi *516*
Digital Archives and Political Legacies: Witll the Obama Corpus Stand the Test of Time?
Losh, Elizabeth *518*
Improving Named-Entity Recognition on Inscriptions on *ukiyo-e* prints: Towards a 'Distant Viewing' in Art History
Machotka, Ewa; Chatzipanagiotou, Marita; Pavlopoulos, John *519*

The Expert in the Loop: Developing a Provenance Linked Open Data Management Platform
Mariani, Fabio *520*
Pose Clustering for Martial Arts Action Recognition: the case studies of Kata and Tai Chi
Marsocci, Valerio; Lastilla, Lorenzo *522*
Differences in Ideologies: Identifying Relevant Topics in Ecuadorian Presidential Speeches from 2007 – 2022
Meneses, Luis *524*
Exploring the Correlations Between Picasso's Artworks and his Personal Contacts
Meneses, Luis; Mallen, Enrique *525*
Architectura Sinica and Collaborative Digital Database Development
Miller, Tracy; Benda, Yuh-Fen; Zhuge, Jing; Zuo, Lala *526*
Multimodal AI support of source criticism in the humanities – work in progress
Muenster, Sander; Bruschke, Jonas; Hoppe, Stephan; Maiwald, Ferdinand; Niebling, Florian; Pattee, Aaron; Utescher, Ronja; Zarriess, Sina *527*
Semiotically Unbound: People, Pedagogies and Dialogues - Beginning to do Digital Humanities at Shanghai University.
Murphy, Orla; Xiao, Shuang *529*
VedaWeb 2.0: Towards a Collaborative Workspace for Indo-Aryan Texts
Neuefeind, Claes; Kölligan, Daniel; Reinöhl, Uta; Sahle, Patrick *530*
Quantifying Representations of Asian Identity in 21st-century Anglophone Fiction for Young Readers
Nomura, Nichole Misako; Dombrowski, Quinn *532*
An International Perspective on Creating an Army of Hacker-Scholars
Öhman, Emily Sofi *533*
Perspectives on the Future of Digital Editions & Publishing
O'Sullivan, James; Pidd, Michael; Murphy, Órla; Wessels, Bridgette *535*
Towards a Global Analysis of Changes in Shape over Time based on Digitised Artefacts: The East-Asian Perspective
Pala, Giovanni; Costiner, Lisandra; Liu, Yidan; Wang, Shuofei *536*
A Method to Automatically Georeference and Estimate the Coastline Precision of Digital Historical Maps
Pala, Giovanni Maria *538*
Modeling Chinese Contemporary Calligraphy: the WRITE Dataset
Pasqual, Valentina; Bisceglia, Marta R.; Iezzi, Adriana; Merenda, Martina; Tomasi, Francesca *539*
Extracting and providing online access to annotated and semantically enriched historical data.The AGODA project
Puren, Marie Anna; Vernus, Pierre; Pellet, Aurélien; Bourgeois, Nicolas *541*
Traven between the impostors. Preliminary considerations on an authorship verification case
Rebora, Simone; Salgaro, Massimo *544*
Callimachus: A tool for the study of the formal contents of ancient Greek papyri
Riaño Rufilanchas, Daniel *547*
Polyphemus, a lexical database of the Ancient Greek papyri, and the Madrid Wordlist of Ancient Greek
Riaño Rufilanchas, Daniel *547*
Close Reading: An Interactive Educational System for Learning How to Read Poetry
Risha, Zak; Ma, Rongqian *549*
Asia News in the 16th and 17th Centuries: Catalogue and Digital Library
Rojo-Mejuto, Natalia; Garrobo-Peral, Manuel *550*
Mining and Modeling Spaces and Places for Literary History as Linked Open Data
Röttgermann, Julia; Hinzmann, Maria; Dietz, Katharina; Gebhard, Henning; Klee, Anne; Konstanciak, Johanna; Schöch, Christof; Steffes, Moritz *552*
The Silences in Archives: A Case Study of Annual Reports from Archives at the National Centre for Biological Sciences, India
Roy, Dibyadyuti; Taparia, Kanishtha; Srinivasan, Venkat *556*
The benefits of increasing the digital availability of Alsatian theater
Ruiz Fabo, Pablo; Werner, Carole; Bernhard, Delphine *557*
Modeling Music for Musicologists: A Linked Open Data Approach
Saccomano, Mark; Shibata, Elisabete; Lewis, David; Hankinson, Andrew; Page, Kevin *560*
Detecting Latent Textual Bias with Topic Modeling and Sentiment Analysis
Sanders, Ashley *561*

Representing scholarly statements in ontologies for data management: The case of musicology
Sanfilippo, Emilio M.; Freedman, Richard *563*
Visualising Women's Lives : A Feminist Approach to Distant Reading
Schreibman, Susan; Barget, Monika *564*
The Agency of Brokerage: Reading Positioning and Power in Character Networks
Selisker, Scott *566*
Genre Classification in English Poetry with Lexical and Prosodic Features
Shang, Wenyi; Underwood, Ted *567*
Structural Balance in the Historical Political Networks of China
Shang, Wenyi; Chen, Song *569*
Developing a text readability system for Sesotho based on classical readability metrics
Sibeko, Johannes; Van Zaanen, Menno *571*
Rule-based Speaker Identification for Speech, Thought and Writing in German Literary Texts
Sluyter-Gäthje, Henny *573*
Semantic Data Lakes for Knowledge Extraction in the Humanities: A Case Study on Bernard Berenson's Network of Acquaintances
Spinaci, Gianmarco; Grillo, Remo; Klic, Lukas; Bonora, Paolo *575*
Commercial crowdsourcing in digital humanities: prospects and ethical issues
Suviranta, Rosa; Hiippala, Tuomo *576*
Geographic analysis of published guidebooks and personal diaries on the diversity of city image in the Edo period
SUZUKI, Chikahiko; KITAMOTO, Asanobu *578*
Digital Humanities in the Wild: Bringing Humanistic Pedagogy to Open Source
Tagliaferri, Lisa *580*
Integrating the Japanese Archaeological Dataset into the ARIADNEplus Data Infrastructure
Takata, Yuichi; Yanase, Peter; Niccolucci, Franco *580*
Queer Coding the Audio Archive: Linked Data and the Lesbian Organization of Toronto (LOOT) Oral History Tapes
Tayler, Felicity; Crompton, Constance *581*
The Japanese Small World of Words. Investigating meaning through a large-scale crowdsourcing study of word associations.
Telegina, Maria; De Deyne, Simon; Joyce, Terry; Miyao, Yusuke *582*
Metadata, Data, and Datasets: An Exercise in Excising the Web Archive for Public Consumption
Thomas, Grace; Dooley, Chase *584*
Infrastructural Sovereignty in the Black Atlantic
Thorat, Dhanashree *585*
Visualizing Rare Books: A Report on Manicule
Trettien, Whitney *586*
"Sweete Flowers and Odoriferous Beds of Spice": Sensory Mining Techniques to Trace Olfactory Orientalism
Tullett, William; Menini, Stefano; Leemans, Inger *587*
The Concept of Nature in German Romanticism: An Approximation
Uglanova, Inna; Gius, Evelyn *588*
Reading *The Unknown* with a Network Mapping Device: Graph Data Visualization for Hyperfiction Works
Viehhauser, Gabriel; Schlesinger, Claus-Michael; Hein, Pascal; Blessing, Andre; Ulrich, Mona *591*
Spatial accessibility of China railway transportation network in the first half of the 20th century
Wang, Changsong; Duan, Yunxin *593*
Modelling Time Ontology in Ancient Chinese Texts
Wang, Linxu; Tong, Wei; Wang, Jun *593*
Quantitative understanding of the evolution of Seo Jeong-Ju's poetic world: Keywords, topics, and sentiments.
Wang, Sungpil; Park, Juyong *596*
Affective Writing in Ming Dynasty *Huaben* Stories — A Topic Modeling Study
Wang, Yiwen; Kurzynski, Maciej *597*
Curuinsi Project: A lexical database for preserving Tikuna language
Wicht, Bertil; Picca, Davide *597*
Learn-STATIC: Building Fundamental Digital Skills in the Humanities Classroom
Wikle, Olivia M.; Williamson, Evan Peter; Becker, Devin; Thornhill, Kate; Hayden, Gabriele *599*

Teaching Basic Buddhism Using a Chatbot: Evaluation and Comparison
Wong, Kwong-Cheong; Chan, Andrew Marcus; Law, Shun-Man; Tse, Devi *600*
From shape to culture: a computational method to extract and study the shape of vases
Yang, Yuchen; Han, Zhitong *602*
Regularization of kinship relations to enrich the family network analysis: Case study on China Biographical Database
Yuan, Yiguo; Li, Bin; Lu, Xuehui; Feng, Minxuan *605*
Dynamic social network tracking in literary texts
Van Zaanen, Menno *606*
Project Overview: Resources and Applications for Detecting and Classifying Polarized and Hate Speech in Arabic Social Media
Zaghouani, Wajdi *607*
How to Critically Utilise Wikidata – A Systematic Review of Wikidata in DH Projects
Zhao, Fudie *608*
Multimedia Retrieval of Historical Materials
Zhu, Jieyong; Nishimura, Taichi; Goto, Makoto; Mori, Shinsuke *610*

Electronic Posters

The Pianolatron: Enabling Web-Based Interactive Performances of Digitized Player Piano Rolls
Abraham, Vijoy; Arul, Kumaran; Barth, George; Broadwell, Peter; Wiles, Simon *614*
Digital City: Developing Digital Guerilla Tactics for the Urban Environment
Altin, Ersin *615*
Revitalizing a South Asian language with Unicode: The case of Sunuwar in Nepal
Anderson, Deborah {Debbie}; Sunuwar, Dev Kumar *616*
RELEVEN: Re-evaluating the Eleventh Century through Linked Events and Entities
Andrews, Tara Lee; Ebel, Carla; Richards, Nina; Prajda, Katalin; Rózsa, Márton; Anđelović, Aleksandar; Read, Lewis *618*
Visualization of annotation system of the Theravada Buddhist literature
Aono, Michihiko *619*
A Study on the Accuracy of OCR-based and NLP-based detection of Japanese Text in the HathiTrust Extracted Features v2.0 Dataset
Bainbridge, David; Hilbing, Genna; Jiang, Ming; Hu, Yuerong; Layne-Worthey, Glen; Downie, J Stephen *620*
Creative Flows: Artistic Inspiration in/through/with Katherine Dunham's Transnational Circulation
Bench, Harmony; Elswit, Kate; Jimenez-Mavillard, Antonio; Uzor, Tia-Monique *621*
"The Great Wall of China" and "Hara-kiri and Feuilleton" or how to search in a digital edition of satirical texts with a focus on the theme of "an oriental fantasy especially oriented about the orient".
Biber, Hanno *623*
Computational Literary Studies Infrastructure (CLS INFRA): a project to connect people, data, tools, and methods
Birkholz, Julie; Börner, Ingo; Chambers, Sally; Charvat, Vera; Cinková, Silvie; Dejaeghere, Tess; Dudar, Julia; Ďurčo, Matej; Eder, Maciej; Edmond, Jennifer; Fileva, Evgeniia; Fischer, Frank; Heiden, Serge; Křen, Michal; Kunda, Bartłomiej; Mrugalski, Michał; Murphy, Ciara; Odebrecht, Carolin; Raciti, Marco; Ros, Salvador; Schöch, Christof; Šeļa, Artjoms; Tasovac, Toma; Tonra, Justin; Tóth-Czifra, Erzsébet; Trilcke, Peer; Van Dalen-Oskam, Karina; Van Rossum, Lisanne *624*
What's new about HuNI?
Burrows, Toby; Verhoeven, Deb *627*
"Es war einmal …" – First Sentences in Literature: A German-Language Reference Corpus
Busch, Anna; Roeder, Torsten *628*
MIV17: a database for 17th-century manuscript culture
Crespi, Serena Carlamaria *631*
Establishing a Code Review Community for DH
Damerow, Julia; Sutton Koeser, Rebecca; Gao, Andrew; Vogl, Malte; Zandbank, Itay; Tharsen, Jeffrey; Casties, Robert; Westerling, Kalle; Carver, Jeffrey *633*

DHTech - An ADHO Special Interest Group
Damerow, Julia; Vogl, Malte; Casties, Robert; Gao, Andrew; Sutton Koeser, Rebecca; Tharsen, Jeffrey; Zandbank, Itay *634*
Fiction, Data: Distant Reading of the Hebrew Novel
Dekel, Yael *635*
Replicating The Riddle of Literary Quality: The litRiddle package for R
Eder, Maciej; Lensink, Saskia; Van Zundert, Joris; Van Dalen-Oskam, Karina *636*
Dehmel digital – Algorithmic-driven indexing of historical letters
Flüh, Marie; Bläß, Sandra *637*
Quantitative Perspectives on European Baroque Drama: Towards a Network Theory-oriented Analysis
Giovannini, Luca *639*
'Double blind' graph data analysis: a pedagogical experiment to discuss the intersubjectivity of network interpretation
Grandjean, Martin; Jacomy, Mathieu *641*
Finding Ortese's Voice for Ferrante Fans: A Stylometric Study of *Neapolitan Chronicles*
Haggin, Patience *642*
Analysis of Exhibition Composition Using Co-occurrence Network Analysis
Hara, Shoko; Ohmukai, Ikki; Nagasaki, Kiyonori; Takagi, Soichiro *645*
Development of datasets of the Hachidaishū and tools for the understanding of the characteristics and historical evolution of classical Japanese poetic vocabulary.
Hodošček, Bor; Yamamoto, Hilofumi *647*
Development of a *Devanāgarī* Optical Character Recognition (OCR) System
Kato, Takahiro; Tomonari, Yūki; Taniguchi, Chikamitsu; Osawa, Tomejiro; Fujimaki, Satoshi; Okada, Takashi; Hashimoto, Emi *648*
Analysis and Exploration of Supernatural Fanfictions from the Platform Archive of Our Own
Kleindienst, Nina; Schmidt, Thomas; Wolff, Christian *649*
Répertoire des Écritures Numériques : archiving and qualifying electronic literature
Lescouet, Emmanuelle; Vitali-Rosati, Marcello *653*
The Vectorian API – A Research Framework for Semantic Textual Similarity (STS) Searches
Liebl, Bernhard; Burghardt, Manuel *654*
A Unicode Input Support Tool for Searching Chinese Characters by Components and Stroke Number
Liu, Guanwei; Nakamura, Satoru; Yamada, Taizo *656*
Centering the Marginalized: Scholar-Curated Worksets from the HathiTrust Digital Library
Magni, Isabella; Worthey, Glen C.; Graham, Maryemma; Walsh, John A.; Downie, J. Stephen; Dubnicek, Ryan C. *658*
Responding to Tibetan Diversity: Rebuilding the Mandala Scholarly Content Management System
Mapp, Rennie; Shinozaki, Yuji; Gunn, Stan *660*
Extraction and Automatic Generation of Characters' Attributes in Contemporary Japanese Entertainment Works
Murai, Hajime; Toyosawa, Shuuhei; Shiratori, Takayuki; Yoshida, Takumi; Nakamura, Shougo; Saito, Yuuri; Ishikawa, Kazuki; Nemoto, Sakura; Iwasaki, Junya; Ohta, Shoki; Ohba, Arisa; Fukumoto, Takaki *661*
Building a Knowledge Base for Data-Driven Historical Information Research Infrastructure and Its Application with Historical Painting Materials
Nakamura, Satoru; Suda, Makiko; Kuroshima, Satoru; Inoue, Satoshi; Yamada, Taizo *663*
Introducing MPCD – Middle Persian Corpus and Dictionary
Neuefeind, Claes; Mondaca, Francisco; Eide, Øyvind; Colditz, Iris; Jügel, Thomas; Rezania, Kianoosh; Cantera, Alberto; Emanuel, Chagai *665*
Do we have to limit our research question by the tool to be used? The iLCM as an example of freely extensible research software for text-based research tasks in the humanities
Niekler, Andreas; Kahmann, Christian *667*
Application for visualizing and analyzing the historical network with context-centric model
Ogawa, Jun; Nakamura, Satoru; Nagasaki, Kiyonori; Ohmukai, Ikki *668*
Considerations for the TEI encoding of Sino-Japanese glossed materials
Okada, Kazuhiro *670*

Handwritten Text Recognition and Palimpsest Analysis for Medieval Greek Manuscripts
Okada, Takashi; Miyagawa, So; Kawazu, Kosei; Ishii, Tatsuya; Oka, Toshio; Fujimaki, Satoshi; Osawa, Tomejiro; Shiki, Yoko; Maehara, Noriko; Ishibashi, Keiichi; Sunada, Kyosuke; Nakanishi, Yasuhito *671*
Uncovering the Black Fantastic: Piloting Text Similarity Methods for Finding "Lost" Genre Fiction in HathiTrust (Poster)
Parulian, Nikolaus Nova; Dubnicek, Ryan; Layne-Worthey, Glen; Williams, Seretha; West-White, Clarissa; Magni, Isabella; Downie, J. Stephen *674*
CREMMALab project
Pinche, Ariane *676*
Buddhist Murals of Kucha on the Northern Silk Road. An Approach to Semi-Automated Annotation
Radisch, Erik *678*
Pragmatic Research Data Management in the Humanities: Dark and Cold Archiving at the Data Center for the Humanities
Rau, Felix; Helling, Patrick; Barabucci, Gioele *679*
Modeling the Multivalent Perspectives of US Immigrant Narrative
Rodrigues, Elizabeth Sarah *681*
Spatio-Temporal Analysis of the Dutch East India Company Archive through Paper Watermarks
Sakamoto, Shouji *682*
CanAICompose? A web-based tool for deep music generation
Schlör, Daniel; Hotho, Andreas *684*
Representing and Modeling Cultural Relevance in Corpora for Historical Analysis
Schroeter, Julian *686*
Developing a Comprehensive Application for Digital Transformation of Historical Materials
Shibutani, Ayako; Nakamura, Satoru; Yamada, Taizo; Yanbe, Koki *687*
Un nouveau partenariat sur les éditions critiques en contexte numérique
Sinatra, Michael; Vitali-Rosati, Marcello; Chateau-Dutier, Emmanuel *689*
HTR2CritEd: A Semi-Automatic Pipeline to Produce a Critical Digital Edition of Literary Texts with Multiple Witnesses out of Text Created through Handwritten Text Recognition
Stoekl Ben Ezra, Daniel; Lapin, Hayim; Brown-DeVost, Bronson; Jablonski, Pawel *690*
Acceptable/Unacceptable/In-between Sentences in Japanese: An Experimental Study on Long-Distance Numeral Quantifiers
Suzuki, Kazunori; Hirano, Michiru; Yamamoto, Hilofumi *691*
Characterizing playing style with speed deviation
Takahashi, Mai; Kobayashi, Michikazu; Ohmukai, Ikki *692*
Extracting clichés: Typify slanderous expressions against the confessions in the #MeToo movement
Takedomi, Yuka; Suda, Towa; Kurita, Kazuhiro; Kobayashi, Ryota; Matsuda, Tomohiro; Uno, Takeaki *695*
Automatic matching method of historical event text with its corresponding thematic maps developed for the application of the ShiJi Spatio-Temporal Information Platform
Tsai, Jung-Yi; Pai, Pi-Ling; Liao, Hsiung-Ming; Chen, You-Jun; Tsai, Richard Tzong-Han; Fan, I-Chun *696*
Using Automated Textual Analysis to Study Concepts of Identity and Difference in First-Person Narrative
Wang, Yadi; Cole, Camille Lyans; Fields, Sam E.; Saelid, Daniel P.; Chen, Annie T. *697*
Issues on Text Encoding of an East Asian Literature
Wang, Yifan; Nagasaki, Kiyonori; Shimoda, Masahiro *698*
Distance Reading Mary McCleod Bethune and the Black Fantastic
Williams, Seretha D.; West-White, Clarissa; Kizer, Ianna; Dickey, Tierany; Albury, Lauren *700*
Humanities Data Inquiry: A Community of Practice Exploring Data Issues in the Humanities and Heritage Research
Woods, Nathan; Bordalejo, Barbara; O'Donnell, Daniel *701*
Distant Reading of the German Coalition Deal
Zylla, Michael; Haider, Thomas Nikolaus *703*

Pre-Conference Workshops and Tutorials

Visual Analysis of Printed Illustrations using Computer Vision

Bergel, Giles Edward
giles.bergel@eng.ox.ac.uk
University of Oxford, United Kingdom

Dutta, Abhishek
adutta@robots.ox.ac.uk
University of Oxford, United Kingdom

This half-day tutorial will provide a practical and theoretical introduction to the computer vision applied to illustrations in various domains. Participants will learn how to make image collections searchable by means of free, open-source tools developed by Oxford's Visual Geometry Group for extracting, matching, comparing and classifying illustrations.

Participants will gain a practical and theoretical understanding of the state of the art in computer vision applied to illustrations. They will learn how to make image collections searchable by means of a modular image processing pipeline composed of free and open-source tools. Participants will learn how to apply, integrate and extend the software tools and processing pipeline to their own images; how visual search and analysis can scale to many millions of images; and will learn how computer vision can provide a deeper understanding of the visual content of image collections. Both the tools and the datasets are based on real-world research projects involving Oxford's Visual Geometry Group and collaborators in the digital humanities and cultural heritage fields.

Relevance to Digital Humanities Audiences

Researchers in many disciplines allied to the digital humanities are interested in the graphical content of books and such other forms of documents as periodicals, posters and pamphlets. While researchers already have many tools for extracting and processing text from documents, there are fewer options for the computational analysis of their visual elements – despite the fundamental importance of non-textual elements in printed communications.

This half-day tutorial is designed to address the needs of such researchers. The tutorial will present a processing pipeline for printed illustrations that are based on the following four open source software applications developed by the VGG based on over a decade of collaboration with different academic disciplines and industrial sectors:

1. **Illustration Detection** using a pre-trained object detector model that has been retrained to automatically detect printed illustrations in early printed books. It has been successfully applied to detect a broad range of printed illustrations (e.g. Spanish Chapbooks). It will be taught in conjunction with the List Annotator (LISA) tool, which is used to review and refine the automatically detected illustrations. The tutorial will show how domain experts can readily use LISA to define regions of interest, and refine the detector by adding missed detections.
2. **Visual image search and grouping** capability is provided by the VGG Image Search Engine (VISE) software which allows visual search of a large collection of images (e.g. a million image) using image (or image regions) as search queries within a graphical interface. VISE is based on features that are robust to different image transformations like rotation, scaling, translation, and shear. Furthermore, VISE uses features extracted from different regions of an illustration which enables search using a part of an illustration. This is useful for identifying damaged illustrations (e.g. due to torn book pages) or illustrations that have been modified in certain ways.
3. **Image Comparison** software allows researchers to finely and forensically investigate the difference between two illustrations which appear similar, but on closer comparison can be seen to have fine differences.
4. **Visual classification** using VGG Image Classifier (VIC). This software incorporates an ImageNet trained model, which can be readily retrained using either local images or images retrieved with user-defined keywords (e.g. ship) via online image search engines (e.g. Google, Bing, etc.). VIC software uses this knowledge to classify and find images in a dataset with content that semantically matches the search keyword.

Participants in the tutorial will step through these applications using the case study data, which will demonstrate both the relevance of these methods for specific use-cases and their general applicability. While the focus of the tutorial is on technical methods in computer vision, it will also cover critical and operational issues such as data capture and cleanup; bias in training data; user experience; and good practice in research reproducibility, software citation and accreditation of invisible labour – issues that digital humanists have a strong interest in foregrounding.

Target Audience

The target audience includes

- Early-career researchers in the humanities wishing to develop their skills.
- Established humanities academics with knowledge of computational methods, not necessarily including computer vision
- Research software engineers based in digital humanities centres or projects
- Academic support staff and research facilitators in digital humanities centres or projects.
- Museum, library and other cultural heritage professionals

The tutorial will be open to all-comers: no prior knowledge of computer vision or programming experience is assumed, but the tutorial will also support technically capable users. The hands-on portion can be followed either through Web demos hosted by VGG, or by user-installable software.

Navigating and Processing Data from the TEI with XSLT

Beshero-Bondar, Elisa Eileen
eeb4@psu.edu
Penn State Erie, The Behrend College, United States of America

Scholger, Martina
martina.scholger@uni-graz.at
Centre for Information Modelling - Austrian Centre for Digital Humanities, University of Graz, Austria

Nagasaki, Kiyonori
nagasaki@dhii.jp
International Institute for Digital Humanities, Japan

Knowing how to navigate and explore data in your encoding can be an important way to learn how to work with TEI and XML generally. This workshop is designed for people who have some experience with markup languages and seek to learn more about how to work with digital scholarly editions as a basis for analysis and research. We seek to raise awareness of the long-term sustainability of XML and TEI and the tool stack designed to process it. A little working knowledge of the query language XPath and the transformation language XSLT can help reduce reliance on software, packages, and plugins that may become obsolete without warning. Further, XSLT's functional programming can serve as a way of articulating research questions around a document data model expressed in XML.

The emphasis of our workshop is “pull-processing”, that is, extracting data and metadata from markup documents for analysis rather than providing the reading view of a digital scholarly edition. Markup in documents supplies structures and contexts that are particularly useful for processing data beyond what we can do with so-called “plain text”. Our workshop will teach the pull-processing of data from XML/TEI with simple, reusable XSLT templates to represent in TSV/CSV and HTML tables/charts.

We will begin by sharing and reviewing together several markup-based projects that visualize data, and we will first study the encoding of these projects. Featured encoding will include projects developed by the workshop instructors and their students, drawn from student projects in the Newtfire organization (https://newtfire.org) , from the East Asian/Japanese SIG of the TEI consortium (https://github.com/TEI-EAJ/) and the SAT Daizokyo Database project (https://21dzk.l.u-tokyo.ac.jp/SAT/index_en.html), and projects in the GAMS (https://gams.uni-graz.at) repository. A sampling of project markup will provide a basis for us to review XML elements and attributes and give participants an opportunity to refresh their understanding of the XML tree structure.

A particularly interesting challenge of our workshop is to share XML documents written in multiple languages. We will also work with source documents composed in languages unfamiliar to the European and North American members of the instructional team, with guidance from our Japanese colleague in order to show that the code we write is transferable to multiple projects across language and cultural borders. Workshop instructors will collaborate on preparing source materials to establish an international foundation for this workshop. This will help us to explore how XSLT works in an international communication and processing context, as long as we are clear on what we wish to process and the significance of the data.

We will demonstrate some basic navigation and calculation functions with XPath, before proceeding to show how XPath is applied in XSLT templates to address specific nodes that hold data of interest for visualization in statistical processing programs and simple online tools, where the structure of the output data is transferable to many different online calculation programs and amenable to statistical processing. During the workshop we will produce some structured documents: HTML lists and tables as well as plain text tabulated data (CSV or TSV files). The XSLT that we supply and that we write together in the workshop will be carefully documented with an interest to assist

participants with revising and adapting the code to their own projects. We also hope to process some participant-supplied XML before, during, and after the workshop.

In a half-day workshop we cannot teach everything that XPath and XSLT can do. However, we can share some commonly used methods to do pull processing (XPath axes, especially descendant::), simple arithmetic (count(), sum(), avg(), multiplication and division operators), output into TSV (CSV) text formats for processing. Making an HTML list and table would be desirable. If we have time, we will model simple transformations of calculated data from XML to SVG to draw a bar or line graph.

Outline

- Review and refresh understanding of XML tree structures, working with XML and TEI projects in multiple languages that explore and analyze data pulled from structured documents.
- Orientation to XPath, working with source documents provided by the instructors.
- Teach basic XSLT to produce simple outputs that curate data and show how these can be used in simple charts and graphs.

Participation and requirements

If you wonder or worry about how to process or share data from your markup, this is a good workshop for you! Familiarity with the TEI in an introductory context is desirable, but participants may also be relatively new to markup technologies. No programming experience is necessary, but some experience with markup is useful. We will provide some introduction to TEI (why TEI is useful for projects that care about data) but we will not have time to do a full introduction of the TEI in the interest of emphasizing simple data processing with existing data. People may find themselves gaining a fresh perspective on the TEI on the basis of what they learn in this workshop.

Our past experience as instructors has shown us that to guarantee a high-quality experience, the instructor team must take great care to monitor and assist our participants. To ensure the instructor team can provide assistance and work with everyone, this workshop is limited to a number of 25 participants. There is no additional charge beyond conference expenses for workshop participants. Instructors will be in contact with participants well in advance of the workshop to provide software installation instructions and helpful resources before we meet, and we will provide means to remain in contact to assist participants after the workshop ends.

Workshop instructors

Elisa Beshero-Bondar, PhD

Program Chair of Digital Media, Arts, and Technology | Professor of Digital Humanities | Director of the Digital Humanities Lab at Penn State Erie, The Behrend College

An active member of the Text Encoding Initiative (TEI), Dr. Beshero-Bondar has been serving since 2016 on the TEI Technical Council, an eleven-member international committee that supervises amendments to the TEI Guidelines. She has been teaching since the 1990s, and began teaching markup languages and XML stack processing almost as soon as she began learning them in the 2010s. Before moving to direct the DIGIT program at Penn State Erie, she was Director of Pitt-Greensburg's Center for the Digital Text. Her research often involves working with archival letters and manuscripts, and has led her to study what early 19th-century poets and dramatists understood about human physiology and electricity, cultural first contact on Pacific islands, and the mutiny on the HMS Bounty. She is the architect of the Digital Mitford Project and other digital research projects involving TEI XML to build editions and prepare structured analyses of variants and collocations in texts. Find her on GitHub at https://github.com/ebeshero and her projects on the site named for her pet firebelly newt at https://newtfire.org.

Dr. Martina Scholger

Centre for Information Modelling - Austrian Centre for Digital Humanities, University of Graz

Martina Scholger has a PhD in Digital Humanities and holds a Senior Scientist position at the Centre for Information Modelling – Austrian Centre for Digital Humanities at the University of Graz. Her main research field is digital scholarly editing, the application of digital methods and semantic technologies to humanities' source material, and text mining. In addition to teaching data modelling, text encoding and X-technologies, her work at the centre involves the conceptual design, development and implementation of numerous cooperation projects in the field of digital humanities (see http://gams.uni-graz.at). She has been an elected member of the TEI Technical Council since 2016, where she is currently its chair, and a member of the Institute for Documentology and Scholarly Editing since 2014. She has been teaching at a number of Summer Schools and workshops in the context of digital scholarly editing, e.g. at the Digital Humanities at Oxford Summer School and Schools organised by the Institute for Documentology and Scholarly Editing (IDE).

Kiyonori Nagasaki, Ph.D.

Senior Fellow at the International Institute for Digital Humanities in Tokyo.

His main research interest is in the development of digital frameworks for collaboration in Buddhist studies.

He is also engaging in investigation into the significance of digital methodology in Humanities and in promotion of DH activities in Japan. He has been participating in a number of Digital Humanities projects conducted at several institutions in Japan and abroad such as the University of Tokyo, Kyoto University, Osaka University, the National Diet Library, the National Museum of Ethnology, the National Institute of Japanese Language and Linguistics, the University of Tsukuba and the University of Hamburg. His activities also include postgraduate education in DH at the University of Tokyo as well as administrative tasks at several scholarly societies including Japanese Association for Digital Humanities, and the Japanese Association of Indian and Buddhist Studies.

Equity, Diversity and Inclusion for Digital Humanists

Bordalejo, Barbara
barbara.bordalejo@usask.ca
Humanities Innovation Lab, University of Lethbridge, Canada

O'Donnell, Daniel
daniel.odonnell@uleth.ca
Humanities Innovation Lab, University of Lethbridge, Canada

Woods, Nathan
nathan.woods@uleth.ca
Humanities Innovation Lab, University of Lethbridge, Canada

This workshop seeks to create awareness of equity, diversity, and cultural differences and present strategies for successful inclusion and the development of fairer environments. These factors are critical within the Digital Humanities because of its emphasis on collaborative work, bringing together different individuals. Moreover, our work has shown that in Digital Humanities, where there could be many different voices, most of the power and prestige remain centralized in the Global-North, in Anglophone countries (Although the community has been making efforts to raise awareness via different statements on inclusion (DHSI Statement on Ethics and Inclusion http://www.dhsi.org/events.php ; ADHO Code of Conduct http://adho.org/administration/conference-coordinating-program-committee/adho-conference-code-conduct ; EADH Diversity and Inclusivity http://eadh.org/about/diversity-and-inclusivity

ACH Statement after the 2016 election https://ach.org/activities/advocacy/ach-statement-in-the-aftermath-of-the-2016-election/ ; CSDH/SCHN Inclusivity and Diversity Statement

https://csdh-schn.org/inclusivity-and-diversity-statement/), these efforts have become part of what we call "the diversity paradox" where success in recognizing there is a problem might often be construed as an achieved solution [See Ahmed 2017, 103]) .

This is not surprising considering the general state of academia as a mostly male, primarily white environment (Johnsrud and Des Jarlais 1994, Towsend 2013). For this reason, it's essential to work with individuals to foster a richer environment, change behaviours, and challenge prejudices.

Reflecting on these matters also allows us to expand the horizons of our limited perspectives. An article published in the Scientific American, "How Diversity Makes Us Smarter," states:

> Decades of research by organizational scientists, psychologists, sociologists, economists and demographers show that socially diverse groups (that is, those with a diversity of race, ethnicity, gender and sexual orientation) are more innovative than homogeneous groups. (Phillips 2014)

In DH, creativity and innovation are essential for the development of new approaches or new applications of traditional methodologies. O'Donnell has argued that diversity is a core value in Digital Humanities:

> "Diversity"—in the sense of access to as wide a possible range of experiences, contexts, and purposes in the computational context of or application of computation to the study of problems in the Humanities, *particularly as this is represented by the lived experiences of different demographic groups*—is, in fact, *more* important than "Quality," especially if "Quality" is determined using methods that encourage the reinscription of already dominant forms of research and experience. (O'Donnell, 2019).

The first version of this workshop was commissioned by Karina Van Dalen-Oskam in 2016 to be delivered to ADHO's steering committee members. This workshop has subsequently been delivered as part of the DH conference on four occasions:

Montreal 2017
Mexico City 2018
Utrecht 2019
Online 2020

We have also delivered versions of this workshop at the Winter Institute for Digital Humanities in India and

developed a new intercultural communication section which was integrated as part of the 2020 version.
Building off this development, in 2019, we unveiled a new game designed to promote the goals of this workshop. These experiences have shown the workshop topic is a moving target as community norms adjust and change (thus, for example, we have seen an increasing interest in the diversity of gender expression in our workshops even over the last six years). Our experience led us to modify and update the workshop's content and activities. We have added a new section on technocolonialism, epistemic alienation, and cognitive justice.

Contents:

1. Diversity, Implicit Bias and Cultural Cloning

Different concepts of diversity, with particular emphasis on cultural and contextual differences. Notions of implicit bias (the unconscious and automatic reflex causing us to pass judgment on others) and Cultural Cloning (the tendency for replication of hiring committees where people of the same ethnicity, ability, gender, etc, hire people like themselves [Essed & Goldberg 2002, Essed 2004]).

The impact of Implicit Bias on teachers and educators, hiring committees and other bodies making decisions about others (journal editors, grant evaluators). We carry out exercises on implicit bias awareness, something which has been shown to have

2. Privilege

To work on the concept of privilege, we have developed a game that we use to illustrate the many factors shaping our lives. Particularly in the case of privileged individuals, it is not easy to locate and categorize their instances of privilege. In collaboration with the Huygens Institute (The Netherlands), we have developed a digital version of the privilege game as a training tool during our workshops (https://privilege.huc.knaw.nl). The game can be played in group or solo modes, and the questions can be modified for particular sets of circumstances.

3. COVID and Technocolonialism

In this section, we explore the impact of technocolonialism (Mboa Nkoudou 2020) during the times of COVID. This follows Mboa's work on identifying epistemic alienation and seeking cognitive justice as proposed by Visvanathan (Visvanathan n.d.) and their application to DH.

4. Intersectionality (in contraposition to the notion of kyriarchy)

When two or more oppressive systems overlap and negatively impact an individual, we discuss intersectionality (Crenshaw 1991, Essed 1994, Essed & Goldberg 2002 Risam 2005). This core concept allows us to explore how privilege or lack thereof impacts individuals directly.

5. Intercultural Communication

A practical section on intercultural communication illustrates the many possible interpretations of specific situations and how these are shaped by individual experiences.

6. Inclusion solutions

Through guided group activities, we facilitate the identification of "inclusion solutions" that participants can implement or propose in their institutions, projects, or other work environments.

At the end of this workshop, we aim to guide participants to reflect on the benefits of a more diverse collaborating and working environment by questioning our preconceived notions of sameness as an ideal.

The workshop is directed at anyone interested in understanding diversity in digital humanities and creating a welcoming and inclusive DH environment. Conference organizers, leaders in the field, and those who often form part of hiring committees are invited to participate. Everyone is welcome to attend, but we particularly encourage the participation of people who are in privileged positions in academia, GLAM, or similar environments.

The workshop is led by Barbara Bordalejo (barbara.bordalejo@uleth.ca) and Daniel Paul O'Donnell (daniel.odonnell@uleth.ca) with the assistance of Nathan Woods (nathan.woods@uleth.ca), Humanities Innovation Lab, University of Lethbridge, A840 University Hall, 4401 University Drive, Lethbridge, Alberta T1K 3M4, +1 403 3292377.

It requires a data projector with audio, as well as internet access.

The activities are self-financing; we do not require participants to pay registration and strongly believe they should not be charged for this particular workshop.

This workshop version will be delivered in four, with capacity for no more than sixty people.

Bibliography

Ahmed, S. (2017). *Living a Feminist Life*. Durham: Duke University Press Books.

Bordalejo, Barbara. (2018). Minority Report: The Myth of Equality in the Digital Humanities. In *Bodies of Information: Intersectional Feminism and Digital Humanities*, edited by Elizabeth Losh and Jacqueline Wernimont, 320–43. Debates in the Digital Humanities. Minneapolis, MN: University of Minnesota Press.

———. (2019). Walking Alone Online: Intersectional Violence on the Internet. In *Intersectionality in Digital Humanities*, edited by Barbara Bordalejo and Roopika Risam. Amsterdam: ARC Humanities. https://doi.org/10.5281/zenodo.2567517 .

Bordalejo, Barbara, Karina Van Dalen-Oskam, Folgert Karsdorp, Daniel Paul O'Donnell, and Basten Stokhuysen. (2019). *Check Your Privilege* (version 1). Amsterdam: Huygens Institute. https://privilege.huc.knaw.nl/.

Bordalejo, Barbara, and Daniel Paul O´Donnell. (2018). Diversity and Collaboration in Digital Humanities: A Workshop. July 13. https://doi.org/10.5281/zenodo.1311946.

Crenshaw, K. (1991). Mapping the margins: intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43.6: 1241-1299.

Essed, P. (1994). Making and breaking ethnic boundaries: women's studies, diversity, and racism. *Women's Studies Quarterly*, 22.3/4: 232-249.

———. (2000). Dilemmas in leadership: women of colour in the Academy. *Ethnic and Racial Studies*, 23.5: 888-904.

———. (2004). Cloning amongst professors: normativities and imagined homogeneities. *NORA - Nordic Journal of Feminist and Gender Research*, 12.2: 113-122.

Essed, P. and Goldberg, D. T. (2002). Cloning cultures: the social injustices of sameness. *Ethnic and Racial Studies*, 25.6: 1066-1082.

Eve, Martin Paul, Cameron Neylon, Daniel Paul O'Donnell, Samuel Moore, Robert Gadie, Victoria Odeniyi, and Shahina Parvin. (2021). *Reading Peer Review*. 1st ed. Cambridge University Press. https://doi.org/10.1017/9781108783521.

Gold, M. (2012). *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press.

Johnsrud, L. and Des Jarlais, C. D. (1994). Barriers to tenures for women and minorities. *The Review of Higher Education*, 17.4: 335-353.

Kim, Dorothy, and Jesse Stommel, eds. (2018). *Disrupting the Digital Humanities*. Goleta, CA: Punctum Books.

Koh, Adeline. (2014). Niceness, Building, and Opening the Genealogy of the Digital Humanities: Beyond the Social Contract of Humanities Computing. *Differences* 25 (1): 93–106. https://doi.org/10.1215/10407391-2420015.

O'Donnell, D. P. (2020). "All Along the Watchtower: Intersectional Diversity as a Core Intellectual Value in the Digital Humanities." In Bordalejo, B. and Risam, R. (eds), *Intersectionality in Digital Humanities*. Amsterdam: Arc Humanities Press.

Phillips, K. W. (2014). How diversity makes us smarter. *Scientific American*. https://www.scientificamerican.com/article/how-diversity-makes-us-smarter/ 1014-42.

Towsend, R. B. (2013). Gender and success in academia: more from the historian's career path's survey. In *Perspectives in History*. http://www.historians.org/publications-and-directories/perspectives-on-history/january-2013/gender-and-success-in-academia

Risam, R. (2015). Beyond the margins: intersectionality and Digital Humanities." *DHQ*, 9.4. http://www.digitalhumanities.org/dhq/vol/9/2/000208/000208.html

———. (2018) New Digital Worlds. Chicago: Northwestern University Press.

Risam, Roopika, and Barbara Bordalejo. (2019). *Intersectionality in Digital Humanities*. Amsterdam: Arc Humanities Press.

Visvanathan, Shiv. n.d. "The Search for Cognitive Justice." Accessed January 22, 2021. http://www.india-seminar.com/2009/597/597_shiv_visvanathan.htm.

Introduction to DraCor – Programmable Corpora for Digital Drama Analysis

Boerner, Ingo
ingo.boerner@uni-potsdam.de
University of Potsdam, Germany

Fischer, Frank
fr.fischer@fu-berlin.de
Freie Universität Berlin, Germany

Milling, Carsten
cmil@hashtable.de
University of Potsdam, Germany

Trilcke, Peer
trilcke@uni-potsdam.de
University of Potsdam, Germany

Sluyter-Gäthje, Henny

sluytergaeth@uni-potsdam.de
University of Potsdam, Germany

Aim of the workshop

In the half-day workshop, DraCor (https://dracor.org), an open platform for researching plays in different languages, will be introduced using practical examples from digital drama analysis. At the center of DraCor are so-called 'Programmable Corpora'. By this we mean infrastructurally research-oriented, open, extensible, Linked-Open-Data-friendly full-text corpora, which should make it possible to address diverse research questions from the field of digital literary studies in a low-threshold way using corpora in a data-based, traceable, and reproducible way (Fischer et al. 2019).

The workshop aims at people who

- work or would like to work with literary texts and in particular with dramas and would like to create their own corpora for this purpose or reuse already existing corpora;
- want to learn methods of digital drama analysis (network analysis, stylometry) or want to try them out on the basis of the Programmable Corpora approach;
- are interested in the possibilities of researching literary texts using Linked Open Data (LOD).

There will be a presentation of the concept of 'Programmable Corpora' as well as a demonstration of the exemplary implementation in the DraCor platform including a presentation of all components. Hands-on tutorials will give participants a practical introduction to creating and curating their own drama corpora for analysis with DraCor. Another part introduces the use of the DraCor API as well as the Python library PyDraCor by means of practical examples on the methods stylometry and network analysis. The Application Programming Interface (API) allows customized direct access to specific parts of the corpora. The possibilities for cross-corpus queries and inclusion of information from the Linked Open Data cloud using SPARQL will also be explored.

The Concept of Programmable Corpora

The core of DraCor consists of corpora of dramas in eleven languages (German, Russian, French, English, Italian, Swedish, Spanish, Ancient Greek, Alsatian, Latin, Bashkir and Tatar) as well as two additional author corpora (Shakespeare, Calderón), to which the platform offers a variety of possible research accesses: The dramas are encoded as XML files according to the TEI guidelines and are freely available under an open license via GitHub at https://github.com/dracor-org. They can be loaded from there, transformed or enriched by oneself if necessary, and reused for further research with any tools.

In addition to this "classical" modus operandi of corpus-based research, however, DraCor as an open digital ecosystem offers further interfaces and connected tools (network visualizations, Shiny App, Easy Linavis). Fundamental to this is the DraCor REST API (https://dracor.org/doc/api), which provides functions for retrieving data in different formats (TEI, JSON, plaintext, RDF, GEXF, GraphML) as well as some built-in analysis functionalities (e.g. on network metrics). The API can be used to retrieve not only structural and metadata, but also the full texts without further markup, so that methods such as stylometric analysis or topic modeling can be applied without any further intermediate step to remove markup. The DraCor API is documented in the OpenAPI standard and can be used in an interactive documentation implemented using Swagger UI (https://dracor.org/documentation/api) directly from the web browser.

API libraries are available for the Python (PyDraCor: https://github.com/dracor-org/pydracor) and R (rdracor: https://github.com/dracor-org/rdracor) programming languages, which allow the API functionalities to be integrated quickly and adapted to the respective programming language. For complex queries, a SPARQL endpoint (https://dracor.org/sparql) is available on the platform. This allows both cross-corpus and combined queries (federated queries), in which DraCor can be queried simultaneously with other resources available as LOD, such as Wikidata.

Digital Drama Analysis with DraCor

Corpus-based analyses of drama, usually using quantitative methods, have emerged as a distinct subfield of Computational Literary Studies (CLS) in recent years (cf. Willand et al. 2017; Reiter 2021). In this context, the provision of jointly curated and open resources such as DraCor has proven productive also for related disciplines such as computational linguistics (cf. for example Pagel, Reiter 2020).

Methods operating at the word level have focused, for example, on authorship attribution (Schöch 2014) or genre classification with topic modeling (Schöch 2017). Currently, promising reconceptualizations of stylometric measures such as the measure Zeta are being developed and applied (Schöch 2018). Furthermore, on the basis of structurally annotated corpora, targeted analyses of, for example of stage directions can be performed, operating with POS information or semantic fields (Trilcke et al. 2020).

In the field of structural analysis, drama corpora were studied early on using network analytic approaches, beginning with the work of Stiller, Nettle, Dunbar (2003) and continuing, for example, with Moretti (2011). Typological work on, for example, the concept of Small Worlds (Trilcke et al. 2016) stands here alongside approaches to the quantitative classification of figure types (Fischer et al. 2018).

Although semantic technologies are now an integral part of the spectrum of methods in the digital humanities, they have rarely been applied in corpus-based CLS (on prose, for example, Frank and Ivanovic 2018; Dittrich 2017). However, the collection of metadata as Linked Data and the connection to external reference resources, especially Wikidata, allow for far-reaching query possibilities and can be profitably used for the analysis of literary corpora. For example, the DraCor corpus data does not contain detailed information on authors and performance locations. However, since the unique Wikidata identifiers are stored for the individual pieces, this information can be retrieved via federated queries in SPARQL and displayed in various visualization forms, such as a map.

Learning objectives and timeline of the workshop

In the first part of the workshop the concept of 'Programmable Corpora' will be introduced and discussed. Afterwards, the platform DraCor and its components will be introduced, with short practical phases during which the participants can directly try out the presented components and tools. In particular, the different possibilities for the reference and analysis of corpus data will be tested. One focus will be on the use of the API. The API functionalities will be explained with the help of interactive documentation and can be tested extensively by the participants. This will be followed by a short overview of corpus creation and the specifics of TEI encoding as used in DraCor.

The second part of the workshop will consist of work phases in which three topics can be explored in more depth:

(1) corpus creation and curation with DraCor: Participants will delve into TEI coding of dramas through hands-on exercises and learn how to set up a local instance of the platform using Docker, customize it if necessary, and populate it with their own corpora.

(2) Drama analysis with DraCor API and Python: Using Jupyter notebooks with extensively documented Python code, participants will be introduced to methods of digital drama analysis using the DraCor API. The notebooks should also enable participants who have no previous experience in programming with Python to follow the individual analysis steps and to adapt them themselves in the sense of a literate programming approach. The notebooks implement concrete research questions on drama analysis, for example on the literary-historical development of network-analytical measures or on the quantitative dominance of characters.

(3) Drama Analysis with Linked Data: The focus will be on practical analyses made possible from connecting DraCor to the Linked Open Data Cloud. The workshop will provide a brief crash course in the SPARQL query language, followed by joint queries of DraCor and Wikidata and visualization of the results.

Organizational matters

Number of possible participants: 25

The Workshop will be held via Zoom. Software to be installed on local machines (Oxygen XML editor, Docker, ...) will be announced in advance. Materials will be made available on GitHub; Jupyter notebooks will be posted at (https://github.com/dracor-org/dracor-notebooks).

Contributors / Contact details

Ingo Börner (ingo.boerner@uni-potsdam.de) works as a research assistant in the project "CLSInfra" at the University of Potsdam on the further development of DraCor. His work focuses on data modeling and Linked Open Data.

Frank Fischer (fr.fischer@fu-berlin.de) is Professor at the Freie Universität Berlin. His involvement with digital drama analysis goes back to the Digital Literary Network Analysis DLINA project (https://dlina.github.io), from which DraCor emerged.

Peer Trilcke (trilcke@uni-potsdam.de) is Professor of modern German literature at the University of Potsdam. His work focuses on the research-based development of infrastructures for literary corpora and the quantitative analysis of literary texts.

Carsten Milling (cmil@hashtable.de) is a web developer and is responsible for the development of the DraCor platform in the project "CLSInfra" at the University of Potsdam.

Henny Sluyter-Gäthje (sluytergaeth@uni-potsdam.de) is a research assistant at the Chair of 19th Century German Literature at the University of Potsdam. She holds a Master of Science in Cognitive Systems with a focus on computational linguistics and works on algorithmic processing of literary texts.

Funding

DraCor is currently being further developed within the EU Horizon 2020 funded project "CLSInfra" (grant number: 101004984, https://cordis.europa.eu/project/id/101004984).

Bibliography

Dittrich, Andreas (2017): "Intra-Connecting an Exemplary Literary Corpus with Semantic Web Technologies for Exploratory Literary Studies" in: Bański, Piotr et al. (Hg.): *Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP)* 2017. Mannheim: Institut für Deutsche Sprache. https://nbn-resolving.org/urn:nbn:de:bsz:mh39-62441.

Fischer, Frank / Trilcke, Peer / Kittel, Christopher / Milling, Carsten / Skorinkin, Daniil (2018): "To catch a protagonist: Quantitative dominance relations in german-language drama (1730–1930)" in: *Digital Humanities 2018. Conference Abstracts.* Mexico City: El Colegio de México / Universidad Nacional Autónoma de México / Red de Humanidades Digitales 193–201.

Fischer, Frank / Börner, Ingo / Göbel, Mathias / Hechtl, Angelika / Kittel, Christopher / Milling, Carsten / Trilcke, Peer (2019): "Programmable Corpora: Die digitale Literaturwissenschaft zwischen Forschung und Infrastruktur am Beispiel von DraCor " in: *DHd2019: »Digital Humanities: multimedial & multimodal«. Book of Abstracts.* Mainz/Frankfurt a. M.: Johannes Gutenberg Universität Mainz / Goethe Universität Frankfurt, 194–197.

Frank, Andrew / Ivanovic, Christine (2018): "Building Literary Corpora for Computational Literary Analysis – A Prototype to Bridge the Gap between CL and DH" in: Calzolari, Nicoletta et al. (Hg.): *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association.

Moretti, Franco (2011): "Network Theory, Plot Analysis " in: *Stanford Literary Lab Pamphlets* 2. http://litlab.stanford.edu/LiteraryLabPamphlet2.pdf [letzter Zugriff 13.7.2021].

Pagel, Janis / Reiter, Nils (2020): "GerDraCor-Coref: A Coreference Corpus for Dramatic Texts in German" in: *Proceedings of the Language Resources and Evaluation Conference (LREC)*. Marseille 55-64 http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.7.pdf [Letzter Zugriff: 15.7.2021].

Reiter, Nils (2021): "Möglichkeiten Quantitativer Dramenanalyse" in: *Comparatio. Zeitschrift für Vergleichende Literaturwissenschaft* 12(2): 39–52.

Schöch, Christof (2017): "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama" in: *Digital Humanities Quarterly* 11, Nr. 2 http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html [Letzter Zugriff: 15.7.2021].

Schöch, Christof (2018): "Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie" in: Bernhart, Toni et al. (Hg): *Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven*. Berlin: de Gruyter 77–94 doi:10.1515/9783110523300-004.

Schöch, Christof (2014): "Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik" in: Schneider, Lars / Schöch, Christof (Hg.): *Literaturwissenschaft im digitalen Medienwandel. Beihefte zu Phin* 7 http://web.fu-berlin.de/phin/beiheft7/b7t08.pdf [Letzter Zugriff: 15.07.2021].

Stiller, James / Nettle, Daniel / Dunbar, Robin I. M. (2003): "The Small World of Shakespeare's Plays" in: *Human Nature* 14: 397–408.

Trilcke, Peer / Fischer, Frank / Göbel, Mathias / Kampkaspar, Dario / Kittel, Christopher (2016): "Theatre Plays as ›Small Worlds‹? Network Data on the History and Typology of German Drama, 1730-1930" in: *Digital Humanities 2016. Conference Abstracts.* Jagiellonian University & Pedagogical University, Kraków 385-387 https://dh2016.adho.org/abstracts/static/dh2016_abstracts.pdf [Letzter Zugriff: 15.07.2021].

Trilcke, Peer / Kittel, Christopher / Reiter, Nils / Maximova, Daria / Fischer, Frank (2020): "Opening the Stage. A Quantitative Look at Stage Directions in German Drama" in: *Digital Humanities 2020. Conference Abstracts*. Ottawa: University of Ottawa https://dh2020.adho.org/wp-content/uploads/2020/07/337_OpeningtheStageAQuantitativeLookatStageDirectionsinGermanDrama.html [Letzter Zugriff: 15.07.2021].

Willand, Marcus / Trilcke, Peer / Schöch, Christof / Rißler-Pipka, Nanette / Reiter, Nils / Fischer, Frank (2017): "Aktuelle Herausforderungen der Digitalen Dramenanalyse" in: *DHd 2017. Digitale Nachhaltigkeit. Konferenzabstracts*. Bern: Universität Bern 175–180 doi:10.5281/zenodo.3684825.

Writing a multilayered article for the Journal of Digital History

Clavert, Frédéric

frederic.clavert@uni.lu
Université of Luxembourg, Luxembourg

Elisar, Ori
ori.elisar@uni.lu
Université of Luxembourg, Luxembourg

Pfeiffer, Mirjam
mirjam.pfeiffer@uni.lu
Université of Luxembourg, Luxembourg

Guerard, Elisabeth
elisabeth.guerard@uni.lu
Université of Luxembourg, Luxembourg

Overview of the tutorial

The first issue of the *Journal of Digital History* (JDH) went out in October 2021. A joint-venture with De Gruyter Publishingq group, the JDH is open access (no fees for authors) and encourages open data. Its distinctiveness is that it implements the concept of multilayered articles, that is based on a long-lasting history of thinking on the future of academic writings (Darnton). This implementation is based on a specifically designed infrastructure based on open source software, with Jupyter notebooks at its heart. JDH's articles are composed of three layers: a narrative layer (the presentation of the results of a research), an hermeneutics layer (exposing methodologies, uses of digital tools, and code), and a data layer (the dataset itself).

The tutorial will last 4 hours and will be divided into two parts. The first part (Two hours, including a break) will show the attendees how to set up their writing environment (https://journalofdigitalhistory.org/en/guidelines). This includes: installing jupyter notebooks and several extensions (nbextensions, cite2c), linking their writing environment to their Zotero account, setting up a github repository (optional, but strongly recommended).

After the familiarization with the environment, the second part will focus on how to write an article and preview it on the JDH front-end (https://journalofdigitalhistory.org/en/notebook-viewer-form - for this interface, using a github repo is mandatory).

For both parts, it is mandatory that attendees come with their own computer with administrator's rights. It is strongly recommended that they have their own data and have a draft article (whatever the journal they intend to publish in) using this dataset in view.

Learning outcomes

- understanding the concept of multilayered article,
- setting up a writing environment that allows interaction with a dataset through code,
- testing their article in the JDH viewer.

Beyond the use of the Journal of Digital History, we aim at showing concretely an alternative way to publish in the digital era, a topic that is fully belonging to Digital Humanities (Fitzpatrick; Vitali-Rosati et Sinatra).

Timeline

Introduction (5 minutes)
The Journal of Digital History and the concept of multilayered article (10 minutes)

First part

Setting up a writing environment (1) (40 minutes)

- installation of jupyter notebooks / lab
- installation of nbextensions / cite2c

break (10 minutes)
Setting up your writing environment (2) (50 minutes)

- linking your jupyter installation to Zotero through cite2c
- using GitHub (optional)

Overview of the main functionalities of notebooks
break (10 minutes)

Second part

Presenting the guidelines of the JDH (including the specific JDH tagging system, 20 minutes)
Presenting the Jupyter Notebook JDH template (10 minutes)
Creating and syncing your repo on GitHub (10 minutes)
Preparing your bibliography on Zotero (10 minutes)
break (10 minutes)
Writing session (45 minutes)

- Including article preview on the JDH.

Wrap up (15 minutes)

Workshop instructors and leaders

Frédéric Clavert is assistant professor in contemporary history and managing editor of the Journal of Digital History. His research are recently focusing on collective memory and social media, as well as on the changing relationships between historians and their primary sources in the digital era.

Elisabeth Guérard is working on the JDH project as an application developer.

Mirjam Pfeiffer is a User Experience and User Interaction Designer, working full time on the Journal of Digital History.

Target audience

We target researchers, mostly but not exclusively historians, who have an experience in writing code to exploit their data but have not yet found a satisfactory way to expose their methods, tools, code and data.

We expect a number of participants around 10 on site. In November 2021, we held a workshop at the French conference DHNord in Lille that aroused some interest, with a smaller audience than the DH conferences'. Aside from that workshop, we also organised several workshops online (Nebraska Lincoln, University of Sussex). Evaluating how many people will attend online is more difficult.

The JDH's team hopes that this tutorial will give them the occasion to meet and deepen their links to the Asian DH community.

Bibliography

Darnton, R. (1999). "The New Age of the Book". *New York Times*.

Fitzpatrick, K. (2011). Planned Obsolescence: Publishing, Technology, and the Future of the Academy. New York: New York University Press.

Vitali-Rosati, M. and Sinatra, M. eds. *Pratiques de l'édition numérique*, Montréal: Presses de l'Université de Montréal, 2014.

Tutorial on Fuzzy String Matching with DeezyMatch

Coll Ardanuy, Mariona
mcollardanuy@turing.ac.uk
The Alan Turing Institute, United Kingdom

Hosseini, Kasra
khosseini@turing.ac.uk
The Alan Turing Institute, United Kingdom

Nanni, Federico
fnanni@turing.ac.uk
The Alan Turing Institute, United Kingdom

Vitale, Valeria
vvitale@turing.ac.uk
The Alan Turing Institute, United Kingdom

Introduction

Fuzzy string matching is a common challenge of linking data in many digital humanities projects, which often deal with noisy, historical, or non-standard text (Olieman et al., 2017). Named entities (in particular place names) are often present under a variety of forms, which can range from regional spelling differences to cross-linguistic or diachronic variation, sometimes due to a change in the political and cultural context, to lack of standardization, or to a process of linguistic standardization. In working with digitized materials, an additional, artificial layer of variation can occur, introduced by optical character recognition errors (Butler et al., 2017; Coll Ardanuy and Sporleder, 2017; De Wilde and Hengchen, 2017; van Strien et al., 2020).

Several studies have warned of the importance of fuzzy string matching for entity linking, especially in noisy and non-standard text (Coll Ardanuy et al., 2020; De Wilde and Hengchen, 2017; Hachey et al., 2013). However, to date, most entity linking systems rely on either exact or partially overlapping string matching. This is due to the high computation time required by most fuzzy string matching approaches, such as Levenshtein distance (Santos et al., 2018a). In this tutorial, we will introduce DeezyMatch (Hosseini et al., 2020), an open-source, user-friendly Python library for fuzzy string matching and candidate ranking for entity linking that has been developed in the Living with Machines project (https://livingwithmachines.ac.uk/).

DeezyMatch builds and expands on Santos et al. (2018b), an approach to fuzzy string matching that uses a deep learning architecture to classify pairs of toponyms as either potentially referring to the same entity or not. DeezyMatch is a tool that integrates recent deep learning advances, and has been specifically designed to be flexible, user-friendly, and fast, and therefore ready to be used in real entity linking scenarios.

In this tutorial, we will show how DeezyMatch can be used to mitigate the problem of name variation in noisy, historical, or non-standard data. We will show how to create string pair datasets that can be used to train and test a DeezyMatch model, and how DeezyMatch models can be used to retrieve candidate entities from a gazetteer or knowledge base. By way of motivation, we will provide and discuss some real digital humanities examples which require fuzzy string matching and will show how DeezyMatch can be used to tackle them. During our tutorial, we will focus on the following case studies:

- **Case study 1:** We will show how a DeezyMatch model can be created from token-level alignments of OCRed text and their manual corrections. We will use the aligned tokens generated in van Strien et al. (2020) using a corpus of OCRed newspaper texts (from the National Library of Australia Trove digitized newspaper collection) that are aligned with human corrections performed by volunteers (Evershed and Fitch, 2014). We will show how to train a DeezyMatch model that learns OCR transformations from newspaper data and will show how it can be used to find a match for a given OCRed query from a pool of potential candidates from a specific knowledge base.
- **Case study 2:** We will show how to create DeezyMatch models that are trained on name variations of places, which will enable us to find the best entry in a gazetteer, for a given query. As an example, we will show how these models can be used to consolidate data about names of heritage locations in Arabic speaking countries, like in the Heritage Gazetteer of Libya (https://slsgazetteer.org/). Currently, the high level of spelling variation in Arabic placenames (across time and transcriptions) makes it difficult to consolidate data that lies in different archives and collections, which at the moment rely on perfect string matching to find connections. We will show how DeezyMatch can be used to more easily associate a heritage location to a number of variant names, thus improving accuracy of data and metadata, and facilitating alignment with other knowledge bases such as Wikidata or Geonames.

This is a hands-on tutorial: participants will be shown how to train a DeezyMatch model and use it for candidate ranking. We will allocate time at the end for discussion, including how to adapt DeezyMatch to different digital humanities projects in different languages and time periods.

We will build on the experience gained on providing two different tutorials on DeezyMatch in the past:

- December 2020: "Linking and Enriching GeoData through Test and Play: a tutorial on DeezyMatch", as part of the *LinkedPasts conference* (Mariona Coll Ardanuy, Kasra Hosseini, Katherine McDonough, and Federico Nanni), followed by a round table. It was held virtually, and there were around 40 participants. Link to the tutorial: https://github.com/LinkedPasts/LaNC-workshop/tree/main/deezymatch
- July 2021: "Best practices in collaborative coding and on using GitFlow for data science research", as part of the *Digital Humanities & Research Software Engineering virtual summer school*, hosted by the Alan Turing Institute. It was held virtually, and there were 25 participants. The focus wat not so much on fuzzy string matching but on collaborative coding. Link to the tutorial: https://github.com/alan-turing-institute/DH-RSE-Summer-School/tree/main/Day%201/gitflow

Outline

This is a half-day tutorial which will cover the following core content:

- **Part 1: Introduction to DeezyMatch and motivation** [60 min]
  - [10 min] Introduction to fuzzy string matching and entity linking
  - [30 min] Description of case studies and data obtaining and preparation
  - [20 min] Overview of DeezyMatch
- **Part 2: Interactive hands-on session** [1h20 min]
  - [10 min] Demo 1: candidate ranking using a pre-trained model
  - [20 min] Hands-on exercise
  - [10 min] Touch base
  - [10 min] Demo 2 and hands-on session: DeezyMatch training and candidate ranking
  - [20 min] Hands-on exercise
  - [10 min] Touch base
- **Part 3: Discussion and feedback** [40 min]
  - [20 min] How to adapt DeezyMatch for your project
  - [20 min] Questions

Instructors

- **Mariona Coll Ardanuy:** Mariona is a computational linguist at the Alan Turing Institute in the Living with Machines project. Her research interests lay in the intersection between the humanities and language technology.
- **Kasra Hosseini:** Kasra is a Research Data Scientist at The Alan Turing Institute. He is interested in (artificially) intelligent systems, machine learning, and data analysis and visualisation.
- **Federico Nanni:** Federico is a Research Data Scientist at The Alan Turing Institute. He is a historian by training and works exploring the intersections between digital humanities, computational social science, and natural language processing.
- **Valeria Vitale:** Valeria Vitale is a researcher in the field of digital cultural heritage. She works at the Alan Turing Institute as Research Associate on the Machines Reading Maps project.

Target audience

Based on past experience, we believe the number of participants should be 20 at most. Participants should have some experience in programming in Python and running scripts, and ideally be interested in entity linking or fuzzy string matching.

Funding statement

This work was supported by Living with Machines (AHRC grant AH/S01179X/1) and The Alan Turing Institute (EPSRC grant EP/N510129/1). The Living with Machines project, funded by the UK Research and Innovation (UKRI) Strategic Priority Fund, is a multidisciplinary collaboration delivered by the Arts and Humanities Research Council (AHRC), with the Alan Turing Institute, the British Library and the Universities of Cambridge, East Anglia, Exeter, and Queen Mary University of London.

Bibliography

Butler, J. O., Donaldson, C. E., Taylor, J. E., and Gregory, I. N. (2017). Alts, Abbreviations, and AKAs: historical onomastic variation and automated named entity recognition. *Journal of Map & Geography Libraries*, *13*(1), 58-81.

Coll Ardanuy, M., Hosseini, K., McDonough, K., Krause, A., van Strien, D., and Nanni, F. (2020). A deep learning approach to geographical candidate selection through toponym matching. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems* (pp. 385-388).

Coll Ardanuy, M., and Sporleder, C. (2017). Toponym disambiguation in historical documents using semantic and geographic features. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage* (pp. 175-180).

De Wilde, M., and Hengchen, S. (2017). Semantic enrichment of a multilingual archive with linked open data. *Digital Humanities Quarterly*.

Evershed, J., & Fitch, K. (2014). Correcting noisy OCR: Context beats confusion. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage* (pp. 45-51).

Hachey, B., Radford, W., Nothman, J., Honnibal, M., & Curran, J. R. (2013). Evaluating entity linking with Wikipedia. *Artificial intelligence*, *194*, 130-150.

Hosseini, K., Nanni, F., and Coll Ardanuy, M. (2020). DeezyMatch: A flexible deep learning approach to fuzzy string matching. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 62-69).

Olieman, A., Beelen, K., van Lange, M., Kamps, J., and Marx, M. (2017). Good applications for crummy entity linkers? the case of corpus selection in digital humanities. In *Proceedings of the 13th International Conference on Semantic Systems* (pp. 81-88).

Santos, R., Murrieta-Flores, P., and Martins, B. (2018). Learning to combine multiple string similarity metrics for effective toponym matching. *International journal of digital earth*, *11*(9), 913-938.

Santos, R., Murrieta-Flores, P., Calado, P., and Martins, B. (2018). Toponym matching through deep neural networks. *International Journal of Geographical Information Science*, *32*(2), 324-348.

Van Strien, D., Beelen, K., Ardanuy, M. C., Hosseini, K., McGillivray, B., and Colavizza, G. (2020). Assessing the impact of OCR quality on downstream NLP tasks. *Special Session on Artificial Intelligence and Digital Heritage: Challenges and Opportunities*, in *Proceedings of the 12th International Conference on Agents and Artificial Intelligence* (pp. 484-496)

Biographical Data in a Digital World 2022 (BD 2022) Workshop

Daza, Angel

j.a.dazaarevalo@vu.nl
Vrije Universiteit Amsterdam, The Nethelands

Fokkens, Antske
antske.fokkens@vu.nl
Vrije Universiteit Amsterdam, The Nethelands; Eindhoven University of Technology, The Nethelands

Hadden, Richard
richard.hadden@oeaw.ac.at
Austrian Academy of Sciences, Austria

Hyvönen, Eero
eero.hyvonen@aalto.fi
Aalto University, Finland; Helsinki Centre for Digital Humanities (HELDIG), University of Helsinki, Finland

Koho, Mikko
mikko.koho@aalto.fi
Aalto University, Finland

Wandl-Vogt, Eveline
eveline.wandl-vogt@oeaw.ac.at
Ars Electronica Research Institute, Austria; Austrian Academy of Sciences, Austria

Biographical and prosopographical (Verboven et al., 2007) data are invaluable sources for historical research as they provide us with essential information on thousands of historical figures: from the cultural heroes of a nation to the many thousands of other significant, yet lesser-known figures who were influential in domains such as the arts, politics, humanities, or natural sciences. Their historical life paths can provide crucial context for tangible heritage objects, which have been created, owned, or influenced by historical actors, or which depict or refer to them. The events of individual biographies can further be aggregated into (histories of) larger contextual composites: groups (e.g., guilds, family histories), institutions (e.g., art schools, universities, religious orders, political movements, companies) and regional entities (from cities to whole countries).

Computational analysis of biographies has opened up new and interesting research directions (Warren, 2018; Fokkens et al., 2017; Tamper et al., 2022). Individuals share common characteristics that can be relatively easily identified and used for information retrieval and data analysis, such as date of birth, occupation, social networks, and locations. Tools and approaches from the digital humanities can be used for both quantitative analyses of such data and for providing leads for more qualitative research questions. New methods of spatial analysis may result in new research questions.

Workshop Goals

This workshop has two main goals. First, previous events on biographical data and digital humanities have revealed that groups working in this domain have different strengths. Though many groups bring researchers from different domains together, some groups mainly consist of domain experts with a lot of knowledge of history, library studies or literature. Other teams are primarily made up of researchers specialized in automatic analysis, formal modeling or visualization. Bringing these groups together increases insights for both types of groups. Technical experts contribute insights on best practices to deal with typical challenges of the domain. Domain experts on what analyses they need, which potential technological challenges are problematic for their research questions and which are not. Second, there is keen interest in cross-national research on biographical information. Most resources are national resources and though many are open, sharing information also means understanding each other's data representation, linking related information, etc.

The *Biographical Data Workshop: modeling, sharing and analyzing people's lives* held at Digital Humanities 2016 in Krakow led to valuable steps on both these points. Among others, the first steps were made that ultimately led to active international collaboration (e.g., the InTaVia project: https://intavia.eu/). Our motivation for organizing this follow-up is two-fold. First, the advances of methods and availability of new resources in the last six years leads us to expect that new interactive sessions of knowledge exchange will have the same beneficiary effect as previous events. Second, the workshop and *Biographical Data in a Digital World* conference series were, in terms of participants, rather European centric with some participants from America. It is our hope that a workshop at the current location will strengthen connections with researchers from Asia. In addition to the general increase of the network of expertise in this domain, steps towards connecting resources for Asia would open new horizons for new comparative research.

Workshop Organization

The workshop will be divided into three sessions.

Preliminary schedule

1. (morning) The participating researchers get the chance to present their work. Depending on the number of submissions, this will be done through (short) presentations or in two poster sessions (where we aim to schedule posters of researchers of existing networks in the same session, to increase options for new connections).
2. (afternoon) The session consists of two interactive themed sessions. In the first round, we have technically oriented themes (data linking and sharing, data analysis and data visualization), where technical experts meet with various domain experts. In the second round, sessions are organized around use cases from the humanities, where humanity scholars meet with a team with different technical expertise.
3. (afternoon) Work groups will share their insights from the interactive sessions with the rest of the participants.

Submissions, proceedings, and workshop participation

We solicit abstract submissions for the poster session through its own call for participation. Possibilities of peer-reviewed proceedings based on the workshop will be discussed with participants during the workshop.

The contributions will be submitted through the main conference ConfTool system: https://www.conftool.pro/dh2022/.
At least one author of each paper needs to register to the DH 2022 conference in order to participate in the workshop. The workshop is open for all conference participants. We reach out for close collaboration with the ADHO Geohumanities SIG (https://geohumanities.org/).

Logistics

Important dates

- March 4: Call for abstracts and requests for participation
- April 14: Deadline for abstracts
- April 29: Notification of acceptance (peer-review results)
- July 25/26: The workshop takes place

Intended audience

The target audience of the workshop are researchers working with biographical or prosopographical data. The audience is expected to consist of DH-focused researchers and computer scientists, as well as social science and humanities researchers.

Expected number of participants

Estimating the number of participants physically present is difficult due to COVID, but the previous BD workshops have had dozens of attending participants; there will be more with the planned possibility for online access.

Intended length

One day.

Budget

The workshop is self-financing.

Topics

Topics of interest include, but are not limited to:

- Digitizing and structuring biographical data
- Standards, vocabularies and best practices for processing biographical data
- Biographies and Linked Data
- Crowdsourcing biographical data
- Automatic biography generation
- Using biographical and prosopographical data for quantitative analyses
- Canonization of people and events in history
- Use of big data for biographical research
- Dealing with biographical data in heterogeneous datasets
- Creating and maintaining biographical dictionaries
- Enriching biographies from external sources
- Reconciling persons between biographical dictionaries
- Reconciling names against a biographical dictionary
- Visualizing biographical and prosopographical data
- Network analysis of biographical data
- Biographies and spatial analysis
- Biographies across countries and cultures

Organizers

Angel Daza, Vrije Universiteit Amsterdam (j.a.dazaarevalo@vu.nl). Angel Daza is a computer scientist with expertise in Natural Language Processing. He completed his Ph.D. research at Universität Heidelberg

where he worked on multilingual models for Semantic Role Labeling. He recently joined as a postdoc for the Computational Linguistics & Text Mining Lab at VU and is part of the In/Tangible European Heritage Visual Analysis, Curation & Communication project (InTaVia).

Antske Fokkens, VU University Amsterdam and Eindhoven University of Technology (antske.fokkens@vu.nl) holds a University Research Chair on Computational Linguistic Methods. The chair specifically focuses on methodological aspects of language technologies when they are used in other disciplines (such as history, literature and library studies). She was one of the main researchers in the BiographyNet project and currently leads the Text Mining package of the InTavia project. Together with Historian Serge ter Braake she initiated the Biographical Data in a Digital World conference series and is co-chair of the DARIAH EU working group analyzing and linking biographical data.

Richard Hadden, Austrian Academy of Sciences (richard.hadden@oeaw.ac.at). Richard Hadden is a Digital Humanities researcher. He completed a PhD in textual scholarship and large-scale digitisation and knowledge representation at Maynooth University, Ireland. He joined the Austrian Centre for Digital Humanities and Cultural Heritage in 2020, as a postdoctoral researcher in digital prosopography. He currently works on several prosopographical and biographical projects as the ACDH-CH, including development of the IPIF prosopographical data format.
https://www.oeaw.ac.at/acdh/team/current-team/richard-hadden

Eero Hyvönen, Helsinki Centre for Digital Humanities (HELDIG) and Aalto University (eero.hyvonen@aalto.fi). Eero Hyvönen is professor of semantic media technology in Aalto University and director of the Helsinki Centre for Digital Humanities (HELDIG); he has been involved in developing several biographical and prosopographical systems for DH research, based on Semantic Web technologies, Linked Data, and Artificial Intelligence.
https://seco.cs.aalto.fi/u/eahyvone/

Mikko Koho, Aalto University (mikko.koho@aalto.fi). Mikko Koho is a Staff Scientist at Aalto University, Department of Computer Science. Research focuses on Linked Data, ontologies, and data modelling, as well as data analysis in multidisciplinary Digital Humanities research.
https://seco.cs.aalto.fi/u/mkoho/

Eveline Wandl-Vogt, (eveline.wandl-vogt@oeaw.ac.at) foundress and director of Ars Electronica Research Institute knowledge for humanity, research manager and experimental researcher at Austrian Academy of Sciences, affiliate at metalab at Harvard, foundress and co-chair of DARIAH EU working group analyzing and linking biographical data, network and open innovation facilitator eg for biographical and prosopographical data, is challenging the genre against an art driven and open innovation background.
https://www.linkedin.com/in/evelinewandlvogt

Confirmed Program Committee members

- Paul Arthur, Edith Cowan University, Australia
- Angel Daza, Vrije Universiteit Amsterdam, The Netherlands
- Thierry Declerck, German Research Center for Artificial Intelligence (DFKI), Germany
- Antske Fokkens, Vrije Universiteit Amsterdam, The Netherlands
- Richard Hadden, Austrian Academy of Sciences, Austria
- Eero Hyvönen, University of Helsinki and Aalto University, Finland
- Mikko Koho, Aalto University, Finland
- Lik Hang Tsui, City University of Hong Kong, Hong Kong
- Eveline Wandl-Vogt, Ars Electronica Research Institute and Austrian Academy of Sciences, Austria
- Hongsu Wang, Harvard University, USA

Bibliography

Fokkens, A. S., ter Braake, S., Ockeloen, C. J., Vossen, P. T. J. M., Legêne, S., Schreiber, G. and de Boer, V. (2017). BiographyNet: Extracting Relations between People and Events. In Bernád, Á. Z., Gruber, C. and Kaiser, M. (eds), Europa baut auf Biographien: Aspekte, Bausteine, Normen und Standards für eine europäische Biographik. Wien: New Academic Press, pp. 193-224.

Tamper, M., Leskinen, P., Hyvönen, E., Valjus, R. and Keravuori, K. (2022): Analyzing Biography Collection Historiographically as Linked Data: Case National Biography of Finland. Semantic Web. Accepted for publication.

Verboven, K., Carlier, M. and Dumolyn, J. (2007). A Short Manual to the Art of Prosopography. In Keats-Rohan, K.S.B. (ed), Prosopography Approaches and Applications. A Handbook. Oxford: Occasional Publications of the Unit for Prosopographical Research, pp. 35-70.

Warren, C. N. (2018). Historiography's Two Voices: Data Infrastructure and History at Scale in the Oxford Dictionary of National Biography (ODNB). Journal of Cultural Analytics, 3(1).
https://doi.org/10.22148/16.028 (accessed 28 April 2022).

Workshop: HathiTrust Research Center's Extracted Features 2.0 Dataset

Dubnicek, Ryan
rdubnic2@illinois.edu
HathiTrust Research Center, iSchool, University of Illinois, Champaign, IL USA

Christie, Jennifer
jechri@iu.edu
HathiTrust Research Center, Pervasive Technology Institute / Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN USA

Kudeki, Deren
dkudeki@illinois.edu
HathiTrust Research Center, iSchool, University of Illinois, Champaign, IL USA

Layne-Worthey, Glen
gworthey@illinois.edu
HathiTrust Research Center, iSchool, University of Illinois, Champaign, IL USA

Walsh, John A.
jawalsh@indiana.edu
HathiTrust Research Center, Pervasive Technology Institute / Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN USA

Downie, J. Stephen
jdownie@illinois.edu
HathiTrust Research Center, iSchool, University of Illinois, Champaign, IL USA

Abstract

This workshop will introduce participants to the HathiTrust Research Center's Extracted Features Dataset, and demo new data fields and functionality introduced in the latest version, 2.0. Generated from the over 17 million volumes (over 60% still in copyright) in the HathiTrust Digital Library, the EF 2.0 Dataset supports text and data mining in this corpus while still being distributed as open, restriction-free data. This tutorial will introduce the EF 2.0 Dataset, the key concepts behind its creation, and hands-on research use cases for the Dataset using IPython notebooks.

Index Terms—digital libraries, text and data mining, HathiTrust, HTRC Extracted Features Dataset, collections as data

Introduction

Digital libraries are in a state of flux: as institutional collections have grown, their purposes have been transformed from "preservation and access" into "big data" repositories. "The Santa Barbara Statement on Collections as Data" identifies foundational principles for this new focus of digital libraries, encouraging library providers and researchers to view these vast digital collections as sources of data for computational research (Santa Barbara Statement on Collections as Data, 2019). This approach opens new possibilities of inquiry and discovery, but also carries with it the weight of copyright limitations and implications; questions of bias, completeness and representation in data; and additional requirements for new skills, sophisticated infrastructures, and increased funding from stakeholders at all levels.

Having foreseen some of these issues, the HathiTrust Research Center (HTRC) developed an approach to "non-consumptive research," a framework that supports computational research of digital items in the HathiTrust Digital Library (HTDL) while complying with copyright limitations on data access. Similarly, through its Extracted Features (EF) Dataset, HTRC supports an approach to computational research that seeks to minimize skill and infrastructure requirements while also aligning with the broader considerations mentioned above. First released in 2015 with version 0.2, the EF Dataset blends MARC-derived, volume-level metadata with computationally generated, page-level metadata and data (Capitanu et al., 2015). The EF Dataset is open (available under a Creative Commons Attribution 4.0 International License), and includes data for each volume in the HathiTrust Digital Library (HTDL) in a bag-of-words format, supporting a multitude of computational use cases.

This workshop will focus on introducing the context and application of the EF Dataset version 2.0, for the first time published in a linked-data-compliant format that includes new and updated data fields, a revised structure, and meaningful uniform resource identifiers (URIs) (Jett, et al., 2020). Attendees will learn about the data from which the EF Dataset is generated, the motivation for the dataset's creation, its structure and format, and its potential applications and limitations. Additionally, workshop attendees will work hands-on programmatically with EF 2.0 data using IPython notebooks, and will use the HTRC FeatureReader Python library, developed specifically to

facilitate use of the EF Dataset (Organisciak and Capitanu, 2016).

Workshop Description

Overview

Section In this three-hour workshop, attendees will first be introduced to HathiTrust, the data of the HathiTrust Digital Library, and the HathiTrust Research Center, followed by context and motivation of the Extracted Features Dataset, the specifics features (or "fields") included in the dataset, its format, and a discussion of its potential uses and limitations. Finally, attendees will have a chance to work directly with Python code to analyze HTRC EF data and to use the HTRC FeatureReader Python library. Hands-on work will also present new linked data fields and their potential application for research with the EF Dataset. A more detailed breakdown of the workshops modules follows:

Section 1: Intro to HathiTrust Digital Library and HTRC

- What is HathiTrust?
- What is/isn't the HathiTrust Digital Library?
- What is the HathiTrust Research Center?

Section 2: Context and motivation for the HTRC EF Dataset

- Non-consumptive research
- What is in the data?
- Data models and analysis techniques

Section 3: Ethical considerations of text datasets

- Bias in libraries, datasets, data, and algorithms

Section 4: Getting and Exploring EF data

- Hands on with EF data, Python notebooks and the HTRC FeatureReader library

The workshop will be a mix of presentation, discussion, and hands-on activity, with an emphasis on open discussion. Our discussion on the ethics of dataset construction, big data and algorithmic analysis will highlight work from Katherine Bode (2020), Catherine D'Ignazio and Lauren Klein (2020), Mimi Ọnuọha (2021) and Roopika Risam (2019).

Audience

The workshop is open to all, and is targeted at digital library and digital humanities scholars of all levels, library and information professionals, and anyone interested in computational research using HathiTrust Digital Library data. Attendees will develop an understanding of HathiTrust Digital Library, the HathiTrust Research Center's services, tools and platforms, selected ethical issues associated with digital libraries, dataset and data analysis, and the Extracted Features model, its suitability for various text and data mining applications, and hands-on familiarity with using the dataset for exploratory data analysis. The workshop can accommodate 10 - 40 attendees.

This workshop builds on years of HTRC workshops ranging from general introductions to text and data mining to more advanced work with HTRC's tools, services, and data. This workshop will provide an in-depth focus on the Extracted Features Dataset version 2.0, demo new ways for exploring and using the dataset and also engage with the ethics of datasets, data and algorithms.

The general objectives of this workshop are to introduce the HathiTrust context, motivation for, and development and release of the Extracted Features Dataset, and to familiarize participants with the data format, its potential applications, and the latest additions in the 2.0 version. Topics of instruction and potential discussion will include:

• How does the Extracted Features Dataset help make the HTDL more accessible for text and data mining?

• What is the EF Dataset model and the structure of its files?

• What research use cases or exploratory data analysis can be supported using HTRC EF data, especially using the new features of the 2.0 dataset"?

• What tools are available for working with EF data, and hands-on experience using them in Python notebooks?

Our goal is for attendees to leave this workshop with a general understanding of the utility of derived datasets and to be comfortable beginning exploratory data analysis using the EF Dataset.

In addition to more HTRC-centric learning objectives, hands-on activities will have added bonuses of an introduction to common cultural analytics tasks in Python, and the associated software libraries used for such tasks, including Pandas, NLTK and Gensim.

Instructor Biographical Information

Ryan Dubnicek is a Digital Humanities Specialist with HTRC, where he works on external and internal research support and outreach and education. He has a Master

of Science in Library and Information Science from the University of Illinois at Urbana-Champaign.

Jennifer Christie is an Associate UX Specialist at the HathiTrust Research Center. She is interested in user-centered interaction design and front-end web development. Her research is grounded in qualitative and quantitative assessments of HTRC's user base, as well as assisting with outreach and education efforts to engage with HTRC's growing user community.

The remaining listed authors have contributed to the development of this curriculum and its associated instructional materials and concepts, but will not be part of the instructional team.

Acknowledgements

The curriculum presented in this workshop was developed with support from HathiTrust, University of Illinois, Indiana University, and the Institute of Museum and Library Services, award number RE-00-15-0112-15. Additionally, former Associate Director of Outreach and Education with HTRC, Eleanor Koehl, contributed heavily to the development and design of the teaching materials and activities.

Bibliography

Bode, K. (2020). Why You Can't Model Away Bias. Modern Language Quarterly, 81(1): 95–124 doi: 10.1215/00267929-7933102.

D'Ignazio, C. and Klein, L. (2020). 'What Gets Counted Counts'. Data Feminism https://data-feminism.mitpress.mit.edu/pub/h1w0nbqp/release/3 (accessed 10 December 2021).

Jett, J., Capitanu, B., Kudeki, D., Cole, T., Hu, Y., Organisciak, P., Underwood, T., Dickson Koehl, E., Dubnicek, R. and Downie, J. S. (2020). The HathiTrust Research Center Extracted Features Dataset (2.0) HathiTrust Research Center doi: 10.13012/R2TE-C227. https://wiki.htrc.illinois.edu/pages/viewpage.action?pageId=79069329 (accessed 2 June 2022).

Organisciak, P. and Capitanu, B. (2016). Text Mining in Python through the HTRC Feature Reader. Programming Historian https://programminghistorian.org/en/lessons/text-mining-with-extracted-features (accessed 2 June 2022).

Risam, R. (2019). The Stakes of Postcolonial Digital Humanities. New Digital Worlds. (Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy). Northwestern University Press, pp. 23–46 doi: 10.2307/j.ctv7tq4hg.5. http://www.jstor.org/stable/j.ctv7tq4hg.5 (accessed 10 December 2021).

Partners Always Already Computational - Collections as Data https://collectionsasdata.github.io/partners/ (accessed 30 November 2021).

The Library of Missing Datasets — MIMI ONUOHA MIMI ONUOHA https://mimionuoha.com/the-library-of-missing-datasets (accessed 10 December 2021).

The Santa Barbara Statement on Collections as Data Always Already Computational - Collections as Data https://collectionsasdata.github.io/statement/ (accessed 30 November 2021).

A picture is worth a thousand words: Image analysis for the Digital Humanities

https://tutorial.thousandwords.art

James, Stuart
stuart.james@iit.it
Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia, Italy

Aubry, Mathieu
mathieu.aubry@enpc.fr
LIGM (UMR 8049), École des Ponts ParisTech, France

Van Noord, Nanne
n.j.e.vannoord@uva.nl
Multimedia Analytics Lab (MultiX), University of Amsterdam, Netherlands

Garcia, Noa
noagarcia@ids.osaka-u.ac.jp
Institute for Datability Science (IDS), Osaka University, Japan

Impett, Leonardo
li222@cam.ac.uk
Cambridge Digital Humanities (CDH), University of Cambridge, UK

Introduction

In this tutorial, we look at Computer Vision (CV) approaches developed to investigate Digital Humanities (DH) data and, more specifically, fine-art and cultural

heritage. We will explain what approaches can achieve, how to train and use with a basic understanding of python to be applied to different types of visual data. By breaking down the tutorial into five parts (one per presenter), the tutorial will provide an overview of the research within CV and its current and future application within DH. We additionally attempt to provide some reflections on the use of Asian data and the limitations or challenges. While considering the current extensive narrative within the CV research community on bias in datasets and collections.

Tutorial Content

Part 1: Retrieval and Knowledge Graphs

The use of CV for distant reading in image collections are generally within the setting of retrieval — searching via an example within a collection. To do this, a computational description of the image needs to generate to be then able to compare one image to another. How such a representation is learned is important to provide a powerful retrieval system. While using pre-trained approaches such as neural networks are useful, they fail to bridge the visual difference between photo-realistic images and common art or humanities image collections. In this part, we explore how anyone can train a neural network representation that is specific to their dataset with varying degrees of supervision, and specifically exploiting supervision that can be provided through Knowledge Graphs (or Semantic Web) to enhance the differential power of the representations.

Part 2: Content-based analysis

Most Deep Learning image techniques rely on annotated collections. While these might be available for some well-studied types of documents, they cannot be expected for more specialized studies or sources. Instead, one would have to rely on techniques that do not require training data. This part will discuss several such techniques to establish links between artworks and historical documents, including the use of generic local features, synthetic data, self-supervised learning, and object discovery techniques. In addition, this will include examples of applications for repeated patterns discovery in artwork collections, fine artwork alignment, document images segmentation, historical watermarks recognition, scientific illustration propagation analysis, and unsupervised Optical Character Recognition. In all cases, it will be shown that standard approaches can give useful baseline results when tuned adequately, but that developing dedicated approaches that take into account the specificity of the data and the problem significantly improves the results.

Part 3: Multi-Task Learning

Multi-Task Learning (MTL) is an increasingly prominent paradigm in CV and in Artificial Intelligence in general. It centers around the ability to perform multiple tasks based on a single input. For instance, it is possible to predict for a single image of an artwork when it was made by who and using what materials. Jointly performing these tasks involves specific modeling choices, resulting in clear benefits (robustness, improved performance), but it also has potential downsides (negative interference, increased complexity). In this part, we show when and how we might want to apply MTL, through a number of use cases, as well as an overview of the technical underpinnings. In addition, highlight the possibilities MTL provides for interpretability by shedding light on relations between tasks.

Part 4: Automatic interpretation

In CV, visual arts are often studied from an aesthetics perspective, mainly by analyzing the visual appearance of an art reproduction to infer its attributes (author, year of creation, theme, etc.), its representative elements (objects, people, locations, etc.), or to transfer the style across different images. However, understanding an artistic representation involves mastering complex comprehension processes, such as identifying the socio-political context of the artwork or recognizing the artist's main influences. In this part, we will explore fine-art paintings from both a visual and a language perspective. The aim is to bridge the gap between the visual appearance of an artwork and its underlying meaning by jointly analyzing its aesthetics and semantics. We will explore multimodal techniques for interpreting artworks, and we will show how CV approaches can learn to automatically generate descriptions for fine-art paintings in natural language.

Part 5: Using Computer Vision within humanities research

Most models in CV research are built to solve specific problems with measurable outcomes (often tied to a set of reference datasets): pixelwise segmentation, object detection, image captioning, keypoint detection, etc. With many open-source computer vision models for each kind of task, we have a wide horizon of powerful tools at our disposal: yet most of them don’t easily fit with research

questions in art history, visual culture studies, or the visual humanities more generally.

By dissecting a series of previous projects in this area, this part will look at how researchers have negotiated these connections, including complex and difficult questions of bias, interpretability, and the epistemology of computational results within the humanities (and especially within cultural history). We will look at several methodological modes compatible with the affordances of CV, including image replication, computational iconography, and the study of visual phenomena captured through notational systems.

Tutorial presenters' brief bios

Stuart James, Istituto Italiano di Tecnologia (IIT) & UCL DH

Researcher (Assistant Professor) in Computer Vision at the Istituto Italiano di Tecnologia (IIT). Stuart's research focus is on Visual Reasoning to understand the layout of visual content from Iconography (e.g. Sketches) to 3D Scene understanding. He is a PI on the MEMEX RIA EU H2020 project and Co-PI on the RePAIR EU FET H2020. Stuart has previously held PostDoc positions at IIT, University College London (UCL) and the University of Surrey. Also, at Surrey, Stuart was awarded his PhD in visual information retrieval using sketches. Stuart continues to hold an honorary position at UCL and UCL Digital Humanities.

Mathieu Aubry, École des Ponts ParisTech

Mathieu Aubry is a tenured researcher in the Imagine team of Ecole des Ponts, focussing on Computer Vision and Digital Humanities. He obtained his PhD from ENS in 2015 and his Habilitation (HDR) in 2019. He had a leading role in several digital humanity projects such as the Young Researcher EnHerit ANR project on enhancing heritage image databases. He was a keynote speaker in several venues, including most recently the ACM Multimedia 2021 workshop on Structuring and Understanding of Multimedia heritAge Contents (SUMAC). He is an associated editor for CVIU and was an area chair at numerous CV conferences.

Nanne van Noord, University of Amsterdam

Nanne van Noord is Assistant Professor in the Multimedia Analytics Lab of the University of Amsterdam. His research is focused on the intersection of multimedia analysis and visual culture. He did his PhD at Tilburg University on modeling the artist's style for recognition and reproduction, as part of the NWO Science4arts project Reassessing Vincent van Gogh. He previously worked in The Sensory Moving Image Archive (SEMIA) project, and coordinated the Computer Vision taskforce in the national infrastructure project CLARIAH.

Noa Garcia, Osaka University

Noa Garcia is an Assistant Professor at the Institute for Datability Science at Osaka University. Her research interests lay at the intersection of computer vision, natural language processing, and art. She has been involved in multiple projects related to computer vision for art, with a special focus on language description and interpretation. She moved to Japan in 2018, after completing her Ph.D. in Computer Science at Aston University, United Kingdom.

Leonardo Impett, University of Cambridge

Leonardo Impett is an Assistant Professor of Digital Humanities at the University of Cambridge. His main work has to do with computer vision for the "distant reading" of art history (CS applied to the humanities), and visual studies as a route to understanding computer vision (the humanities applied to CS). He was previously based at Durham University, Harvard University, the Max Planck Institute for Art History, and EPFL.

Schedule

25th July 2022			
JST	**UTC**	**Topic**	
18:00 - 18.05	09:00 - 09:05	Introduction	
18:05 - 18:35	09:05 - 09:35	Part 1	Retrieval and Knowledge Graphs
18:35 - 19:10	09:35 - 10:10	Part 2	Content-based Analysis
Break - 15 Minutes			
19:25 - 20:00	10:25 - 11:00	Part 3	Multi-Task Learning
20:00 - 20:35	11:00 - 11:35	Part 4	Automatic interpretation
Break - 15 Minutes			
20:50 - 21:25	11:50 - 12:25	Part 5	Using Computer Vision within humanities research
21:25 - 21:30	12:25 - 12:30	Close	

Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870743

Data and Algorithms in Critical Aging Studies

Karadkar, Unmil
unmil@karadkar.me
University of Graz, Austria; University of Texas at Austin, USA

Kriebernegg, Ulla
ulla.kriebernegg@uni-graz.at
University of Graz, Austria

Sawchuk, Kim
kim.sawchuk@concordia.ca
University of Graz, Austria

Taipale, Sakari
sakari.taipale@jyu.fi
Concordia University, Canada

Ivan, Loredana
loredana.ivan@comunicare.ro
National University of Political Studies and Public Administration (SNSPA), Bucharest, Romania

Introduction

This workshop will invite participation from scholars who are studying the interplay between aging and digital technologies. Ever-increasing data collection through mobile services, online communications, embedded and worn sensors, social media platforms, and digitization is affecting several aspects of our lives, including aging. The data generated is algorithmically processed by corporate, governmental, individual, and non-governmental actors to shape the experiences, services, and opportunities available to an increasingly aging population. Recognizing that neither data nor algorithms are neutral, their detrimental effects on disadvantaged demographics have been widely studied in humanistic contexts, most famously in "Algorithms of Oppression" (Noble, 2018), which articulates the racial biases in google's search algorithms.

Organized by leading scholars from aging studies (Katz, 2014), communication studies, information sciences, and social sciences, we will engage scholars who use humanistic approaches to research the impact of data and algorithms on the aging population. The significance of aging as an issue is recognized by the World Health Organization, which has formed a global collaboration around this issue titled "the UN Decade of Healthy Aging" (World Health Organization, 2021).

A non-exhaustive list of topics of interest includes:

- impact of digitization, data, and algorithms on the processes and practices of aging
- aging, statistics, and public policies
- data ageism - ethics, harms, and justice
- policies and infrastructure for responsible international and interdisciplinary data sharing
- new methods in interdisciplinary, intersectional scholarship
- representations of ageing and data in literature and media
- humanistic studies of datafication of aging
- agency and autonomy in sensor-enhanced ageing
- digital technology practices, impositions, and appropriation
- community-driven data
- development of datasets to document the changing reality of aging adults
- innovative qualitative, quantitative, and ethnographic approaches to critique data
- theorization of datafication through the lens of aging

Workshop organizers

This workshop is proposed by the PI and some co-applicants of Aging in Data, a Canadian SSHRC funded (number: 895-2021-1020) partnership project consisting of 19 partner organizations and 34 co-applicants and collaborators from 10 countries in Europe, North America, and Australia (Concordia press release, 2021).

Unmil P. Karadkar (unmil.karadkar@uni-graz.at) works as a Scientist in the Center for Interdisciplinary Research on Aging & Care (CIRAC) at the University of Graz, Austria. He situates his work at the intersection of aging studies and research data management, sharing, and reuse. His research contributes to areas such as the design of digital collection interfaces and digital humanities. Unmil also holds a research appointment at the University of Texas at Austin, USA. His research has been funded by the National Science Foundation, Texas General Land Office, USAA, and Andrew W. Mellon Foundation.

Ulla Kriebernegg (ulla.kriebernegg) directs the Center for Interdisciplinary Research on Aging and Care (CIRAC) and is Associate Professor of American Studies at the University of Graz, Austria. Her research focuses on North American literary and cultural studies, Aging and Care

Studies, and Medical Humanities. Her latest book, Putting Age in its Place (forthcoming) deals with the spatiality of care in cultural representations of care homes. Ulla is chair of the Age and Care Research Group Graz and deputy chair of the European Network of Aging Studies (ENAS). She co-edits the Aging Studies book series (Transcript), is Associate Editor of The Gerontologist and member of the GSA's Humanities, Arts, & Cultural Gerontology Panel. She directs the Graz part of SSHRC's "Aging in Data" project.

Kim Sawchuk (kim.sawchuk@concordia.ca) is a Professor in the Department of Communication Studies, holds the Concordia University Research Chair in Mobile Media Studies, and is the Associate Dean of Research and Graduate Studies for the Faculty of Arts and Science at Concordia University, Montreal, Canada. Her research at the intersection between aging and communication technologies challenges lingering ageist assumptions within media studies, where old age and new media are often positioned as incommensurable topics. Kim is a co-founder of the Critical Disability Studies Working Group and has led Ageing, Communication, Technologies (http://www.actproject.ca), a Canadian SSHRC-funded interdisciplinary, international, multi-methodological project that investigates the transformation of experiences ageing with the proliferation of new forms of mediated communications. She is also the PI of the SSHRC-funded Aging in Data partnership project (2021-2028).

Sakari Taipale (sakari.taipale@jyu.fi) is Associate Professor of Social and Public Policy at the Department of Social Sciences and Philosophy, University of Jyväskylä, Finland, and a part-time Senior Research Associate at the Faculty of Social Science, University of Ljubljana, Slovenia. In Jyväskylä, he leads the 'New Technologies, Ageing and Care' research group of the Centre of Excellence in Research on Ageing and Care (CoE AgeCare). Dr Taipale has participated in many European research networks and he has made research and teaching visits to universities in countries such as Italy, Lithuania, Slovenia and Spain.

Loredana Ivan (loredana.ivan@comunicare.ro) is associate professor PhD, at the National University of Political Studies and Public Administration (SNSPA), Bucharest, Romania. She studies technology-mediated interpersonal communication, combining sociological and social psychological approaches. With a background in Sociology, she holds a PhD from the University of Bucharest. She has been Marie Curie Scholar at the University of Groningen, Interuniversity Center for Methodology (ICS), in The Netherlands and visiting researcher at Humboldt University, Berlin, Department of Social and Organizational Psychology. She is also a co-founder of the Research Group in Social Psychology at the Society of Romanian Sociologists (SSR) and the current chair of the European Network on Ageing Studies (ENAS)

Diversity and Inclusion

The organizers come from Europe and North America, are gender-balanced as well as diverse in academic disciplines and career stages. We will strive to achieve similar balance in our presenter pool to the extent possible.

Similar to Europe and North America, digital technology is affecting the experience of aging in countries such as Japan, China, India, and South Korea. We seek to engage the Asian Aging Studies scholarly community in aligning with the theme of the conference as well as to broaden our professional network.

Target audience and expected number of participants

This is our first workshop on this topic and we are unsure what the interest will be. In this light, we have tried to make the topic as broad as possible, while retaining the focus on our core interests. Given the increase in issues related to socio-humanistic studies of aging and the increasing focus on critical studies of computing and data, we anticipate receiving 10 to 20 submissions and accepting 8 to 10 for presentation at the workshop. We are unable to gauge the level of interest in the DH community in attending the workshop without presenting.

Technical requirements - Projection, Internet access

The workshop will not require special technical support. A projector for presentations and an internet connection for participants to showcase web-based materials as well as for remote participation will suffice.

Length and format - Half day

This time frame will allow for adequate exploration for a newly forming community. We are open to conducting a full-day workshop if the reviewers believe that there will be a larger interest than we anticipate.

The workshop will be held in a seminar style, with a keynote, followed by short and long presentations. Individuals may participate in the workshop without presenting. In order to facilitate community-building, the organizers will include an open discussion time or breakout sessions depending upon the audience size and time limitations.

Budget - None

The workshop will have no budgetary requirements. All participants will pay for their attendance. The workshop will not require supplies or other resources that will result in costs.

Covid19 accommodations

While we encourage in-person participation, we recognize that regional Covid19 travel or health conditions may make it difficult for presenters to attend in person. Presenters may present in person or remotely. We will accommodate last minute changes.

Code of conduct

In order to ensure a safe, respectful, and collegial environment at the workshop, the organizers will adopt the ADHO DH conference code of conduct, available at:

https://adho.org/administration/conference-coordinating-program-committee/adho-co nference-code-conduct.

Solicitation of proposals

We will invite proposals for long (up to 750-1,000 words, 20 min) and short (up to 300 words, 10 min) presentations. Long presentations will report on mature work in the area or stake out a position in an area while short presentations will present early results or nascent work.

Program Committee

Submissions will be vetted by a small program committee, which currently consists of the organizers. If necessary, we may include other colleagues to review proposals.

Calls for Proposals

We will advertise the workshop CfP through various DH as well as Aging Studies networks (NANAS - North American Network in Aging Studies, ENAS - European Network in Aging Studies, Socio-GeronTechnology Network). We are not aware of similar aging networks in the Asia-Pacific and welcome pointers from the DH2022 program committee.

Timeline

February 16: Announcement and release of CfP
May 15: Long and short presentation proposals due
May 25: Proposal decision notification
July 25: Workshop at DH 2022

Bibliography

Noble, S. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism.* New York University Press, 256pp. ISBN13: 978-1-4798-4994-9

Katz, S. (2014). What is Age Studies? In *Age Culture Humanities 1(1)* pp. 17-23.

World Health Organization (2021). *The UN Decade of Healthy Aging*. https://www.who.int/initiatives/decade-of-healthy-ageing (accessed 30th Nov. 2021).

Concordia Center for Research on Aging, press release (2021). 2 Concordians are awarded nearly $5m for social sciences and humanities research. https://www.concordia.ca/news/stories/2021/07/07/2-concordians-are-awarded-nearly-5m-for -social-sciences-and-humanities-ressearch.html?c=/research/aging (accessed 30th Nov. 2021).

Literary Text Analysis with Spyral Notebooks, a Notebook Environment Companion to Voyant Tools

Land, Kaylin Catherine
kaylin.land@mail.mcgill.ca
McGill University

Rockwell, Geoffrey
grockwel@ualberta.ca
University of Alberta

MacDonald, Andrew
andrewjames.code@gmail.com
McGill University

Tchoc, Bennett Kuwan
tchoh@ualberta.ca
University of Alberta

Damasah, Elliot
damasah@ualberta.ca
University of Alberta

Digital literary text analysis is increasingly becoming an integral part of literary studies. However, many tools designed for performing such analysis remain inaccessible to researchers without significant coding and computing skills. Voyant Tools was designed in part to address this gap. Spyral Notebooks are an extension of Voyant Tools and allow researchers to expand upon their findings from Voyant in a notebook environment. Unlike other notebook environments, Spyral Notebooks are accessible without downloading any programs or advanced set-up. Spyral Notebooks are available in an entirely online format. To use Spyral Notebooks, one needs only a connection to the Internet. The notebooks are easily adaptable, shareable, and editable.

Spyral is a notebook development environment that is integrated into Voyant Tools. Notebook environments can be thought of as both extensions of traditional research notebooks and as novel tools that integrate documentation, active analysis and presentation of results. At their core, notebooks are made up of three types of blocks or cells that a user can add or delete in a sequence.

1. There are text cells that can contain headings and other text elements found in word processors or browser editors (usually based in HTML) for typing unstructured text. Depending on the notebook environment, the text blocks can be simple or more sophisticated. Spyral Notebooks use HTML for text and offer an in- browser WYSIWYG HTML editor for the text blocks.
2. There are code cells where the user inputs code, be it Python, the Wolfram language used in Mathematica, or JavaScript, which is used in Spyral. The code cells can be run in sequence or individually as you debug your code. Code cells can contain as much or as little code as the user desires.
3. There are output cells which produce the output of the code you input in the associated code cell. It is important to recognize that the output of the code is dependent on what you have instructed the computer to do; that is, it is not a printout of the code cell but the results of running your code. You thus have to instruct the computer to print out the desired results.

In our tutorial we introduce participants to Spyral Notebooks. We illustrate how to create a corpus for textual analysis from Voyant Tools or directly in Spyral Notebooks. After walking through the basic mechanisms for using Spyral Notebooks including saving, editing, and sharing notebooks, we move on to more specific features available in Spyral. Participants will learn how to enhance the capabilities of Voyant and go deeper with their textual analysis using Spyral. Finally we provide participants with several tutorial notebooks designed to highlight some of Spyral's advanced features such as categories for use in sentiment analysis.

Spyral Notebooks are a welcome addition to the field of digital humanities as they provide an accessible notebook environment specifically designed for literary text analysis. Spyral Notebooks are thoughtfully designed to serve researchers with limited coding skills who want to take their analysis from Voyant one step further. We especially envisage Spyral proving useful for digital humanities instructors. Spyral provides a useful platform for student work, allowing students to embed their analysis from Voyant, perform more complex analysis using JavaScript, and annotate their code with their thought processes.

Making 3D-scans more mobile-friendly and increasing online audience reach: Introduction to manual retopology

Leelasorn, Angel
angel.leelasorn@anu.edu.au
Australian National University, Australia

Digital heritage has become increasingly visible in recent years and the popularity of digital museums has been on the rise, in part fuelled by the emergence of COVID-19 (Davis, 2020). Digital heritage scholars and museums alike now have an online presence and many are sharing 3D models of artefacts or historical sites created using photogrammetry, LiDAR, or other 3D-scanning methods on online platforms, e.g. Sketchfab, to increase public engagement. Digital 3D models have also seen use in digital preservation of damaged historical sites. One example is the digital reconstruction of Shuri Castle in Okinawa, Japan where its real-world counterpart was destroyed in a fire in 2019 (Shiraishi, 2019). Ubisoft, one of the world's major video game companies, has also published a Virtual Reality experience of Notre-Dame de Paris which is available for free on Steam, a video game distribution service, after the fire in 2019.

However, many of the 3D models of scanned historical artefacts or sites published by scholars on online 3D model sharing platforms are often poorly optimised and therefore unsuitable to be viewed on mobile devices because of their large file size. Since the majority of the users of these platforms are mobile users, the museums and scholars are at risk of losing engagement due to users being turned

away from viewing the shared 3D models. The platforms have also implemented additional features exclusive to mobile devices, e.g. Sketchfab's default high polygon count models filter, which filters out models which are deemed too heavy to be run on most portable devices and essentially discourages users from viewing the performance-intensive models.

In addition, there is a tendency for digital heritage scholars to avoid learning the geometry-based 3D modelling pipeline also used by the video game industry due to its steep learning curve, and opt for more feasible methods such as photogrammetry (Rahaman and Champion, 2019). There is also a lack in the fundamental knowledge of performance optimisation for 3D modelling and real-time rendering among digital heritage scholars, as the available online resources are focused on video game art. A clearly-defined and versatile workflow for digitisation is also lacking (Rahaman, Champion and Bekele, 2019).

A solution to this problem is performance optimisation for real-time rendering. Manual retopology is one of the performance optimisation techniques used by the video game industry to reduce the amount of triangles on a 3D model . While photogrammetry software, e.g. Agisoft Metashape, include features such as the decimation tool to reduce polygon count, there are still limitations with the algorithm-based 3D model generation. These limitations primarily exist since the algorithm is typically unable to prioritise potentially crucial detail of a 3D model, unlike the human eyes. It is only possible to algorithmically reduce the polygon count to a certain number before the faces start to become visible when examined in close proximity. The topology of the generated model is also unsuitable to be repurposed in other contexts, e.g. a virtual reconstruction or a Virtual Reality experience. In contrast, even though manual retopology is generally more time-consuming than algorithmic decimation, it allows precision and customisation to the point that the polygon count can be reduced from millions of triangles to merely thousands or even hundreds and drastically reduces the file size while maintaining the quality, fidelity and appearance of a 3D model. Therefore, manual retopology is an essential skill that is often overlooked in the field of digital heritage and is certainly beneficial once incorporated into the workflow.

Bibliography

Davis, B. (2020). In a year when many were stuck indoors, Google says 'virtual museum tours' was among its most popular search terms. *Artnet News*. https://news.artnet.com/art-world/virtual-museum-tours-1930875 (accessed 8 December 2021).

Rahaman, H. and Champion, E. (2019). To 3D or not 3D: choosing a photogrammetry workflow for cultural heritage groups. *Heritage*, *2*(3), pp.1835-1851.

Rahaman, H., Champion, E. and Bekele, M. (2019). From photo to 3D to mixed reality: A complete workflow for cultural heritage visualisation and experience. *Digital Applications in Archaeology and Cultural Heritage*, *13*, p.e00102.

Shiraishi S. (2019). Shuri Castle: Fire destroys 500-year-Old world heritage site in Japan. *BBC News Japan*. https://www.bbc.com/news/world-asia-50244169 (accessed 8 December 2021).

Getting Started with the Advanced Research Consortium

Liebe, Lauren
leliebe@tamu.edu
Texas A&M University, United States of America

Mandell, Laura
mandell@tamu.edu
Texas A&M University, United States of America

Tarpley, Bryan
bptarpley@tamu.edu
Texas A&M University, United States of America

Workshop Description

This workshop will serve as an introduction to the Advanced Research Consortium (ARC), a vibrant community of researchers who peer review and curate digital cultural heritage materials for humanities research.

Since 2004 with the founding of NINES, ARC has the dual goals of providing a vetting community for digital scholarship in particular fields and a technological infrastructure to support dissemination and use of this scholarship. In addition toNINES.org, ARC communities include 18thConnect.org,MESA-medieval.org, Modernist Networks (ModNets.org), and SiRO (Studies in Radicalism Online.org), with others currently organizing in the fields of early American studies and disability studies. The ARC community seeks to develop itself towards greater cultural inclusiveness.

The workshop begins by teaching researchers how to use ARC to enhance their research through searching both

open access and proprietary resources curated by field experts. Because ARC's catalog includes a wide array of resources not generally cataloged by university libraries or other discovery services, using ARC allows for more nuanced searching, particularly for primary materials. Additionally, this portion of the workshop will preview the Corpora Dataset Studio for researchers interested in further developing their own digital projects.

Next, the workshop familiarizes researchers with the process of submitting their own projects or datasets for inclusion in ARC. To contribute to ARC, projects must undergo both content and technical peer review. Once peer review has been successfully passed, metadata about each digital artifact in the project is added to ARC's catalog, which then links back to the project itself. Inclusion in ARC makes independent digital humanities projects more easily discoverable and allows for greater interoperability between different projects and proprietary data from resources such as JSTOR and Adam Matthew. Furthermore, ARC's work with the Linked Infrastructure for Networked Cultural Scholarship (LINCS) project means that each included project will be enhanced through the inclusion of linked open data.

Finally, the workshop will discuss the process for creating new period-specific or thematic research nodes within ARC. Historically, ARC's nodes have been quite broad, encompassing broadly defined periods such as the eighteenth century or modernism. However, in order to facilitate greater collaboration between digital humanities projects, ARC is now developing smaller, more specific nodes to address individual scholarly communities. In order to expand our offerings, ARC seeks proposals for editorial groups who wish to serve scholarly communities with particular interests — e.g., early modern book history, eighteenth-century Caribbean literature, animal studies, Victorian children's literature, etc. This portion of the workshop will include discussion of forming content and technical editorial boards, locating relevant projects, managing peer review, and working with the ARC office to create a user interface.

Workshop Instructors

Dr. Laura Mandell (mandell@tamu.edu) — Laura Mandell is the author of Breaking the Book: Print Humanities in the Digital Age (2015), Misogynous Economies: The Business of Literature in Eighteenth-Century Britain (1999), and, recently, "Gendering Digital Literary History: What Counts for Digital Humanities," in the New Companion to Digital Humanities (Blackwell 2016). She is Project Director of the Poetess Archive, an online scholarly edition and database of women poets, 1750-1900 (http://www.poetessarchive.org), Acquisitions Editor of 18thConnect (http://www.18thConnect.org), and Director of ARC (http://www.ar-c.org), the Advanced Research Consortium overseeing NINES, 18thConnect, and MESA. She spearheaded the Early Modern OCR project or "eMOP" (http://emop.tamu.edu), a project concerned with improving OCR for early modern and 18th-c. texts via high performance and cluster computing, and is currently at work on a text-mining project to discover emergent genders in essays and novels comprising the Feminist Controversy debates in England, 1788-1810.

Dr. Bryan Tarpley (bptarpley@tamu.edu) — Bryan Tarpley is Associate Research Scientist of Critical Infrastructure Studies at the Center of Digital Humanities Research at Texas A&M University. He is also the Associate Director of Technology for the Advanced Research Consortium (ar-c.org). Dr. Tarpley's recent software development includes the Corpora Dataset Studio, to be released open source in 2023.

Dr. Lauren Liebe (leliebe@tamu.edu) — Lauren Liebe is a postdoctoral researcher at Texas A&M University, where she serves as the project manager of the Advanced Research Consortium. Her doctoral research focused on the intertwining of politics and theatrical performance during the English Restoration. She is also the creator of Digital Restoration Drama (restorationdrama.org), a database of TEI-encoded Restoration playtexts.

Description of target audience and expected number of participants

The target audience for this workshop is scholars who are interested in collaborating to form new thematic research nodes to join the Advanced Research Consortium.

This workshop is open to all interested participants.

Syllabus (3 hour workshop)

- What is ARC? (10 minutes)
- Using ARC (45 minutes)
  - Researching via the nodes
  - BigDIVA
  - Corpora
- Contributing to ARC (40 minutes)
  - Submitting for peer review
  - Peer review process
  - Metadata ingestion
  - Linked open data
- Break (15 minutes)
- Developing an ARC Node (50 minutes)
  - Determining scope
  - Directors and editorial boards

- Determining relevant resources for the research environment (digital projects, datasets, journals, library collections, proprietary digital resources)
- Designing your Collex instance
- ARC Office Contributions
  - Conducting Peer Review
  - Negotiating, obtaining, and ingesting metadata from open access projects and vendors
  - Node maintenance, including the solicitation of classroom and exhibit features.
- Q&A session (20 minutes)

Scholarly writing and editing with the text editor Stylo

Mellet, Margot Lise
margot.mellet@umontreal.ca
Chaire de Recherche du Canada sur les écritures numériques, Canada

Fauchié, Antoine
antoine.fauchie@umontreal.ca
Chaire de Recherche du Canada sur les écritures numériques, Canada

Vitali-Rosati, Marcello
marcello.vitali-rosati@umontreal.ca
Chaire de Recherche du Canada sur les écritures numériques, Canada

Presentation of the Stylo-sophy

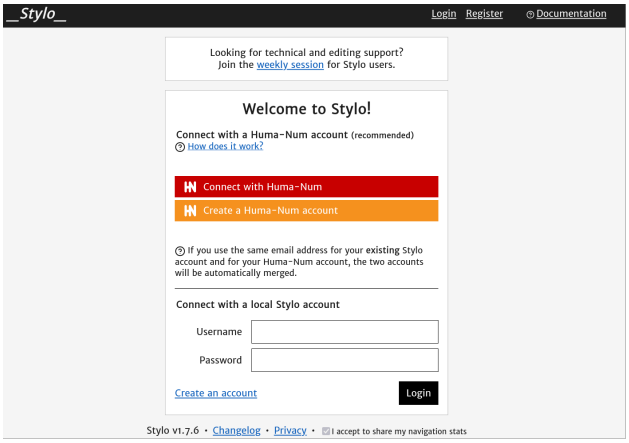


Figure 1.
Login page of Stylo

Stylo is a tool designed to transform the digital workflow of scholarly journals in the humanities and social sciences. As a WYSIWYM (What You See Is What You Mean *)* semantic text editor for the humanities, it aims to improve the academic publishing chain.

In order to put authors back in control of the text (Vitali-Rosati et al, 2020), text structuring is placed at the very beginning of the knowledge production process through the use of semantic tags. Unlike word processing tools such as Microsoft Word, Stylo aims to promote and encourage the use of open standards (notably HTML, Markdown, XML). This is achieved through the implementation of open protocols.

Tool principles

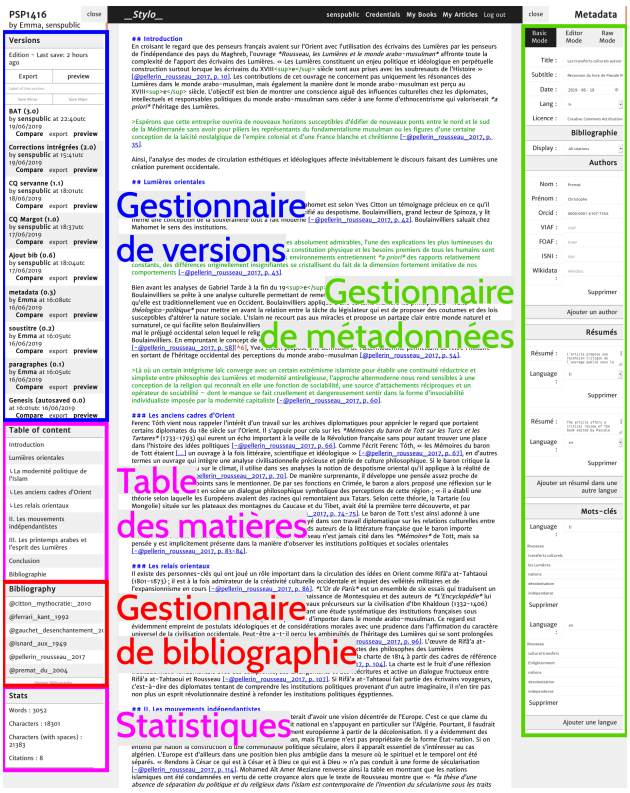


Figure 2.
Stylo editing organisation

Built on modular, low-tech and standard editing tools and formats, such as Markdown, BibTeX, Pandoc, Hypothes.is and Latex, Stylo integrates best practices in writing and publishing on the web into a single interface. Through the implementation of the formats and the conversion technologies already used by the community,

it allows the free flow of documents that are not locked into a particular format. Stylo includes features such as sharing, versioning, change tracking, bibliographic reference management, annotation for revisions, multi-format export, metadata aligned with authorities (LOC, Wikidata, ORCID, ...) and online semantic tagging.

Workshop structure

In this workshop, we will present how Stylo is used daily as a writing and editing tool for a scientific journal, and how it can be used by researchers and students, individually or collectively. The demonstration will include

1. an introduction to the theoretical basis of the tool
2. a presentation of the main editing features (editing the text body, structuring the bibliography, defining the metadata)
3. a presentation of the export options available
4. a presentation of the specific use of Stylo in the context of scholarly journals in the humanities
5. a presentation of current and future developments and implementations of the tool.

Participants will have the opportunity to test and edit their own text in Stylo. The workshop will use existing Stylo documentation: the documentation site and the video tutorial.

Stylo is currently used by a small and growing community, and has been available since 2020 as one of the services offered by the TGIR Huma-Num.

During the workshop, we want to collect feedback on the tool's interface, ergonomics and usability, but also on the scientific writing practices that Stylo could integrate.

Duration: two hours.

Persons in charge of the workshop:

Antoine Fauchié, PHD candidate in the Department World Literature and Languages at the University of Montreal and project manager of the Canada Research Chair on Digital textualities.

Roch Delannay, PHD candidate in the Department World Literature and Languages at the University of Montreal and project assistant of the Canada Research Chair on Digital textualities

Bibliography

Vitali-Rosati, M., Sauret, N., Fauchié, A. and Mellet, M. (2020). Écrire les SHS en environnement numérique. L'éditeur de texte Stylo. *Revue Intelligibilité du Numérique*, 1, http://intelligibilite-numerique.numerev.com/numeros/n-1-2020/18-ecrire-les-shs-en-environnement-numerique-l-editeur-de-texte-stylo (accessed 14 April 2022).

Ugarit: Translation Alignment Technologies for Under-resourced Languages

Workshop presented at DH2022 Tokyo

Palladino, Chiara
chiara.palladino@furman.edu
Furman University, United States of America

Yousef, Tariq
tariq.yosef@uni-leipzig.de
University of Leipzig, Germany

Shamsian, Farnoosh
shamsian@informatik.uni-leipzig.de
University of Leipzig, Germany

Kanagawa, Nadia
nkanagawa@furman.edu
Furman University, United States of America

In this workshop, participants are going to learn the fundamentals of Translation Alignment and learn to use Ugarit (http://ugarit.ialigner.com/), an online environment targeted at the creation of manually aligned datasets in different languages. The goal of the workshop is to introduce participants to an important topic in Digital Humanities, and to expand our community and available datasets by targeting East Asian languages and Japanese in particular.

Translation alignment is one of the most important tasks in Natural Language Processing. It is defined as the comparison of two or more texts in different languages, also called parallel texts or parallel corpora [6][10], by means of automated or semi-automated methods. The result often takes the form of a list of pairs of items, which can be words, sentences, or larger text chunks like paragraphs or documents. The aligned pairs may be one-to-one (one word in the source text corresponds to one word in the translation), but often align as one-to-many, many-to-many, or many-to-one. Each word correspondence may be complete or perfect (with complete overlap between two words), but also possible or incomplete (partial overlap, or

both words being a translation of each other only in certain contexts [4]).

There are numerous methods for automated translation alignment: the most popular ones, such as statistical machine translation, are based on various levels of manually aligned training data [2], although new models are being proposed, such as neural machine translation [1]. However, the alignment of texts in different languages is an exceptionally complex task, especially when considering word-level alignment. It is often difficult to find perfect correspondences across languages that express ideas through different morphosyntactic constructs, with variations in word order, sentence length. In addition, it is notoriously difficult to establish correspondences within wordplays, metaphors, or allusions. For these reasons, manually aligned word pairs are extremely important to establish gold standards, as sources of training data to implement machine translation methods, and for many other purposes, including text mining and creation of dynamic lexica [4][9].

Some modern languages, like English, German, and Chinese, have an impressive infrastructure for managing parallel corpora. However, that is not the case for historical and generally under-resourced languages, such as Classical Arabic, Persian, Latin, Ancient Greek, Gaelic, Cherokee, Georgian, and even for many languages of East Asia, including Japanese, Korean, and Sanskrit. Ugarit (http://ugarit.ialigner.com/) is a web-based environment designed to support the needs of these languages, providing a framework for creating and using manually aligned corpora. During the workshop, we will introduce the tool and illustrate the many ways in which parallel corpora aligned with Ugarit are currently used: these will include pedagogy and language learning, interlinguistic and translation analysis, dynamic visualization, data mining, dynamic lexica, and training of machine translation models [5][7] [8][11]. We will invite the participants to test the tool on their own corpus or with a prepared dataset, to try first-hand the work of translation alignment, and to visualize and investigate the results.

Ugarit currently supports most East Asian languages and alphabets, but there are very few aligned datasets currently available. With this workshop, we want to specifically target the creation of new parallel corpora in Japanese, Chinese and Korean, and gather more feedback and requests from this part of the Digital Humanities community.

Instructors:

Tariq Yousef is a research associate at Leipzig University, working on Computational Linguistics, Textual Alignment, and Data Visualization. He is the Lead developer of Ugarit. Contact: tariq.yosef@uni-leipzig.de .

Chiara Palladino is Assistant Professor of Classics at Furman University. As project partner in Ugarit, she uses the tool in teaching and research and has led multiple workshops and seminars on translation alignment. Her main interest lies in language learning processes with translation alignment. Contact: chiara.palladino@furman.edu .

Farnoosh Shamsian is a PhD candidate at Leipzig University. As a project partner in Ugarit, she uses the tool in teaching and research and has led multiple workshops and seminars on translation alignment. Her main interest lies in digital pedagogy and teaching Greek through digital annotations. Contact: shamsian@informatik.uni-leipzig.de .

Nadia Kanagawa is James B. Duke Assistant Professor of Asian Studies and History at Furman University. She is a Ugarit user who often works with and translates classical Japanese texts in her research on immigrants in the early Japanese state. Contact: nkanagawa@furman.edu .

Bibliography

[1] Bahdanau, D., Cho, K., and Bengio, Y. “Neural Machine Translation by Jointly Learning to Align and Translate.” (2016). ArXiv:1409.0473 [Cs, Stat], May. http://arxiv.org/abs/1409.0473.

[2] Brown, P. F. et al. “A Statistical Approach to Machine Translation.” Computational Linguistics 16.2 (1990): 79-85.

[3] Dagan, I., Church, K., and Gale, W. “Robust Bilingual Word Alignment for Machine Aided Translation.” In Natural Language Processing Using Very Large Corpora, edited by Susan Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann, and David Yarowsky, 209–24. Text, Speech and Language Technology. Dordrecht: Springer Netherlands (1999). https://doi.org/10.1007/978-94-017-2390-9_13.

[4] Graça, J. et al. “Building a Golden Collection of Parallel Multi-Language Word Alignment.” LREC (2008).

[5] Foradi, M. “Confronting Complexity of Babel in a Global and Digital Age. What can you produce and what can you learn when aligning a translation to a language that you have not studied?” DH2019: Digital Humanities Conference, University of Utrecht, July 9-12. Book of Abstracts. 2019. https://dev.clariah.nl/files/dh2019/boa/0611.html;

[6] Kay, M. and Röscheisen, M. “Text-translation alignment.” Computational Linguistics 19.1 (1993): 121-142.

[7] Palladino, C. “Reading Texts in Digital Environments: Applications of Translation Alignment for Classical Language Learning.” Journal of Interactive Technology and Pedagogy, 18 (2020). https://jitp.commons.gc.cuny.edu/reading-texts-in-digital-environments-applications-of-translation-alignment-for-classical-language-learning/

[8] Palladino, C., Yousef, T., and Foradi, M. “Translation alignment for historical language

learning: a case study". Digital Humanities Quarterly 15.3 (2021), http://www.digitalhumanities.org/dhq/vol/15/3/000563/000563.html

[9] Véronis, J. "From the Rosetta Stone to the Information Society." In Parallel Text Processing. Alignment and Use of Translation Corpora, edited by Jean Véronis, 1–24. Springer Science & Business Media, 1999.

[10] Véronis, J. Parallel Text Processing: Alignment and Use of Translation Corpora. Springer Netherlands, Dordrecht-Boston-London (2000).

[11] Yousef, T., and Jänicke, S. "A Survey of Text Alignment Visualization." IEEE Transactions on Visualization and Computer Graphics PP (October 2020): 1–1. https://doi.org/10.1109/TVCG.2020.3028975.

From Concepts to Textual Phenomena and Back: Operationalization in the Digital Humanities

Pichler, Axel
axel.pichler@ts.uni-stuttgart.de
University of Stuttgart, Germany

Krautter, Benjamin
Benjamin.Krautter@uni-koeln.de
University of Cologne, Germany

Pagel, Janis
janis.pagel@uni-koeln.de
University of Cologne, Germany

Andresen, Melanie
melanie.andresen@ims.uni-stuttgart.de
University of Stuttgart, Germany

Abstract

This tutorial addresses one of the central challenges of the digital humanities: the operationalization of theoretical concepts from the humanities for computer-based research.

While humanities scholars primarily work with concepts that often encompass several textual phenomena and furthermore draw on contexts deemed relevant for their interpretation, computer-based work is bound to identifiable phenomena on the textual surface. The resulting discrepancy between theoretical expectations and concrete results needs to be bridged by an adequate operationalization. Thereby, the goal is to develop procedures to trace back theoretical concepts to text surface phenomena, potentially in several sub-steps. Or, in short: to detect and measure instantiations of theoretical concepts.

With our tutorial, we want to focus on this practice and its theoretical backgrounds: On the basis of selected use cases, we will show which challenges arise from the use of computational methods for questions in the humanities and how they can be dealt with. In a practical part, the participants will have the opportunity to work on the operationalization of relevant concepts for exemplary text analyses. For this purpose, we provide Jupyter notebooks for the prepared use cases. Programming skills are not required.

The aim of the tutorial is to raise awareness of the differences between established methods in the humanities and computer-based approaches, to address typical challenges, and to develop approaches for adequately operationalizing theoretical concepts from the humanities. We are convinced that reflecting the underlying assumptions of an operationalization is the only way to ensure that one can then handle the results appropriately.

Target Audience

This tutorial targets all researchers with an interest in reflecting their understanding of operationalization theoretically and practically. Group size between 15 and 25 people would be preferable.

Schedule

(total 4 hours incl. 30 min. break)

1. Introduction and procedure (10 min.)
2. Theoretical part (40 minutes in total)
   - Problem outline
   - Introduction to the use cases
3. Practical part
   - Introduction to the primary texts and tools, distribution of the draft guidelines (10 min.)
   - First practical round (small groups): manual annotation of a phenomenon, parallel extension/ revision of the guidelines, iterative (40-50 min.)
   - - Coffee break (30 min) -
   - Collection of results and discussion of approaches (20 min.)
   - Second practice round (small groups): work on operationalization toolbox, feedback on output file, iterative (40-50 min.)

4. Final discussion: collecting results, discussion of experiences and learning objectives (40 min.)

Tutorial Instructors

Melanie Andresen

melanie.andresen@ims.uni-stuttgart.de
University of Stuttgart
Institut for Natural Language Processing
Pfaffenwaldring 5b
D-70569 Stuttgart

Melanie Andresen is a postdoc researcher at the Institute for Natural Language Processing at the University of Stuttgart. She studied German Linguistics at the University of Hamburg and received her PhD in corpus linguistics there in 2020. She can draw on a lot of experience with the operationalization of questions in the humanities and social sciences from the projects hermA (University of Hamburg) and Q:TRACK (University of Stuttgart, University of Cologne).

Benjamin Krautter

Benjamin.Krautter@uni-koeln.de
University of Cologne
Department for Digital Humanities
Albertus-Magnus-Platz
D-50931 Cologne

Benjamin Krautter is a PhD student at the Department of German Studies at the University of Heidelberg and a member of the Q:TRACK project (Cologne). Among other things, he is working on the operationalization of literary concepts for quantitative drama analysis. In his research he focuses on how to meaningfully combine quantitative and qualitative methods for the analysis and interpretation of literary texts.

Janis Pagel

janis.pagel@uni-koeln.de
University of Cologne
Department for Digital Humanities
Albertus-Magnus-Platz
D-50931 Cologne

Janis Pagel is a PhD student at the Institute for Natural Language Processing at the University of Stuttgart and research associate at the Department for Digital Humanities at the University of Cologne. He studied German studies and linguistics in Bochum, and computational linguistics in Stuttgart and Amsterdam. His research focuses on the application of computational linguistic methods to literary studies and coreference resolution on literary texts.

Axel Pichler

axel.pichler@ts.uni-stuttgart.de
Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung
Pfaffenwaldring 5b
D-70569 Stuttgart

Axel Pichler studied Philosophy and Literary Studies in Vienna and Graz (Austria). Currently, he is working as a postdoc on, among other things, the development and reflection of methods of computer-aided text analysis at the Institute for Machine Language Processing at the University of Stuttgart.

Hands-on Introduction to eScriptorium, an Open-Source Platform for HTR

Stokes, Peter Anthony
peter.stokes@ephe.psl.eu
École Pratique des Hautes Études – Université PSL, France

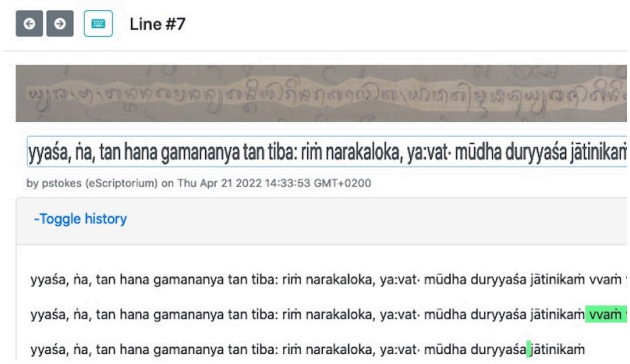
Stökl Ben Ezra, Daniel
Daniel.Stoekl@ephe.psl.eu
École Pratique des Hautes Études – Université PSL, France

The goal of this tutorial is to introduce participants to the principles and hands-on practice of the eScriptorium platform for the automatic and/or manual segmentation and transcription of manuscripts and printed books in a very wide range of languages, writing-systems and complex page-layouts.

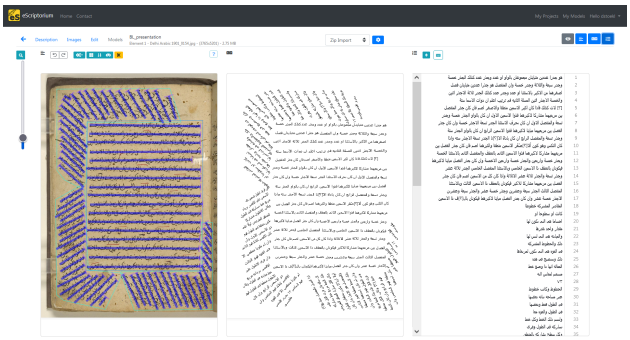
The tutorial will begin with a very brief overview of Handwritten Text Recognition (HTR) and its potential and remaining challenges, particularly when applied to so-called “rare” or historical scripts. HTR has long been a goal – indeed dream – of many in the Digital Humanities and beyond, and this is now becoming realised and increasingly accessible. Relatively general models for automatic transcription can now often achieve character error rates (CER) of less than 1% even for manuscript material, while a CER of 10% or even 20% is sufficient for some types of analysis such as text identification or text-image alignment. Very large quantities of manuscript material are also becoming available, particularly with the increasingly widespread use of IIIF. Machine learning is

also relatively accessible, including computers with GPUs that are sufficiently powerful to treat thousands or even tens of thousands of images of manuscript pages within a reasonable timeframe. This combination is enabling new possibilities that were not feasible only a few years ago. However, most of the existing systems were designed at least initially in European (primarily Anglophone) contexts, and the differences between these and other writing contexts may seem subtle but very often make software unusable in practice.

eScriptorium is Free Open-Source Software [1] designed to facilitate transcription of manuscripts in a wide variety of languages, scripts and complex layouts. The software began as part of the Scripta-PSL project which incorporated experts in dozens of scripts and languages, including Sumerian, Ugaritic, Syriac, Arabic, Hebrew, Classical and Pre-Imperial Chinese, Old and Medieval Japanese, Tibetan, Devanagari, Old Khmer, Pali, Tamil, and many more (see further https://scripta.psl.eu/langues/). It is designed to interact with kraken, another Free Open-Source Software developed by Benjamin Kiessling (of the same institution as the eScriptorium team). [2] Kraken provides a flexible, modular HTR engine that can be run on its own or as the engine behind eScriptorium.



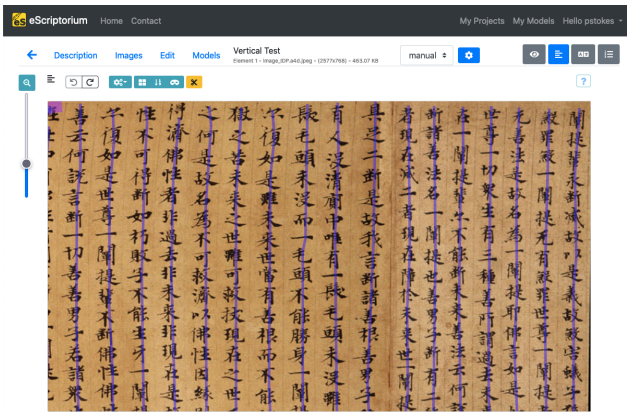
Transcription (transliteration) of Old Javanese palm-leaf books with eScriptorium (courtesy of Marine Schoettel, École Française d'Extrême-Orient)



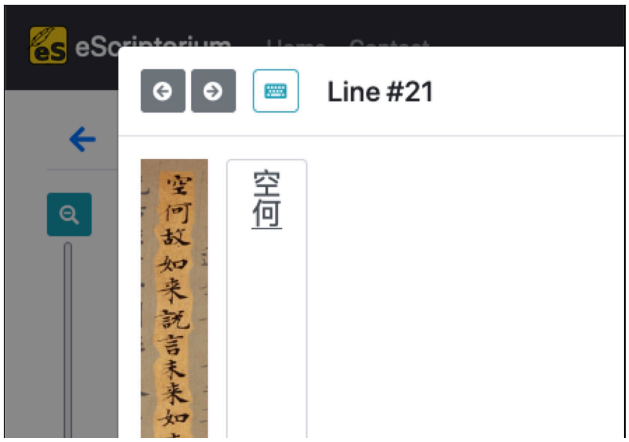
The interface showing a complex page layout: page image with lines and regions (left panel); transcription aligned to the page layout (centre panel); transcription in 'regular' lines (right panel).

Participants will be introduced to eScriptorium and kraken, focussing particularly on the key design decisions that distinguish them from other HTR systems. One is that the system should be as independent as possible of any assumptions about language or script. For instance, writing can be left-to-right, right-to-left, top-to-bottom then right to left (as in Japanese), or top to bottom then left to right (as in Mongolian). The writing can be on a baseline (as in English), from a top-line (as in Hebrew) or in a column (as in Chinese), and the lines can be oriented at any angle, including upside-down and therefore 'reversed' from the normal direction relative to the page image (so upside-down Arabic reads from left to right relative to the page orientation, and so on). Participants are therefore strongly encouraged to bring their own materials in different scripts and languages, and will learn the possibilities and limits of eScriptorium and kraken for these different cases.

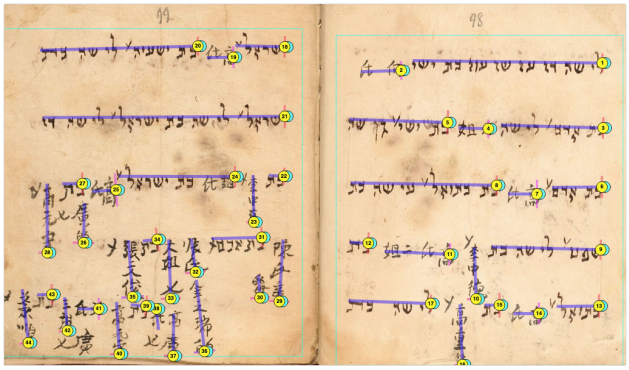
After an introductory session on the principles of HTR and of eScriptorium and kraken in particular, we will go through the different stages of the workflow, always with attention to the challenges of non-European scripts. This involves first, segmentation of pages into regions and lines; second, potentially modifying the order of lines on the page if it has not been correctly identified by the system; and, third, performing the transcription. The first and third of these are currently based on machine learning (Kiessling 2020), and so they both normally involve a process of creating ground truth data for a sample of content, training a model on the basis of this data (most likely on top of an existing one already trained on similar content), then applying this model to new images. This process in turn raises methodological questions, such as which categories of region and line types to use, which standards of transcription to use, how best to share corpora for ground truth, how to manage large character sets or abbreviations, and so on. Participants will be directed to existing work and initiatives on this topic such as SegmOnto, HTR for All, and OCR-D.



Uncorrected results of automatic line detection in a previously unseen scroll (writing in Chinese from top to bottom then right to left).



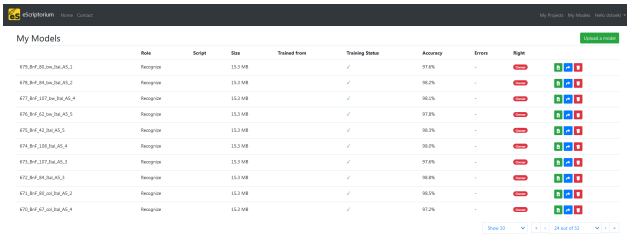
Manually transcribing Chinese (top to bottom then right to left)



The eScriptorium interface showing lines showing mixed Hebrew (right to left) and Chinese (top to bottom or right to left). Numbers indicate the order of reading (detected automatically but correctable manually).

eScriptorium and kraken are also designed to be as open as possible, including that data should be interchangeable and that users should not be locked into any one instance of the framework. Participants will therefore learn how to import and export images, page layout data and transcriptions in a variety of standard formats (plain text, ALTO, PAGE XML; import from IIIF manifests; transformation of export to TEI). We will also briefly discuss the strengths and limitations of these formats, particularly for the wide range of different scripts.

As well as importing and exporting data, users of eScriptorium and kraken are also able (and actively encouraged) to import and export trained models for layout analysis and transcription, including not only sharing with other users in the same instance of eScriptorium but also publishing models on external repositories. A Zenodo community has been established to help share models, and publishing to and from Zenodo is already implemented as part of kraken and will be shortly in eScriptorium. In this way, models can be reused across different instances of the software, and no user is locked into any single instance. This has clear advantages in reducing the human effort in repeating training and ground-truth generation, and also provides a small but not insignificant step towards reducing the energy and environmental impact of machine learning, in that it reduces the need to pointlessly retrain identical models with the same ground truth. This is in contrast to many other systems of machine learning, where some or all of the software may be Open Source, and where the ground truth can sometimes be exported, but the models themselves are locked into a given instance of the system and cannot be exported, and/or the software is not sufficiently open that an entirely new instance can be set up independently.



Model management in eScriptorium, including export, import, sharing and basic version control. Models shown are for manuscripts written in Hebrew

eScriptorium also provides a web API to allow for automated execution of project-specific processes, and so a brief introduction to this will also be provided, depending on time and the interest of participants.

Finally, eScriptorium and kraken are both freely available for anyone to download and install on their own servers, and a number of groups in different countries have already done so to our knowledge. In practice, eScriptorium can run even on a home computer for most tasks, but training on a large corpus requires relatively powerful

computers and servers, and so the last part of the tutorial will be spent discussing practical questions about how to get an instance up and running, how to plan a future project with the software, and any other issues that come up during the session.

Acknowledgements

This work was supported by grants from the European Union (RESILIENCE RI and Vietnamica), Université Paris Sciences et Lettres (Scripta-PSL), the Mellon Foundation (OpenITI), the French Ministry of Higher Education and Research, the French Ministry of Culture (LectauRep) and the Domaine d'intérêt majeur STCN (ManuscriptologIA).

For a full list of contributors to eScriptorium, see the GitLab repository and wiki (links below).

Bibliography

Chagué, A. [no date]. Prendre en main eScriptorium. https://lectaurep.hypotheses.org/documentation/prendre-en-main-escriptorium. English version (transl. Jonathan Allen), https://lectaurep.hypotheses.org/documentation/escriptorium-tutorial-en

eScriptorium[source code]: https://gitlab.com/scripta/escriptorium/

Gabay , S., Camps, J.-B., Pinche, A., and Jahan C. (2021). SegmOnto: common vocabulary and practices for analysing the layout of manuscripts (and more). *16th International Conference on Document Analysis and Recognition (ICDAR 2021).*

HTR United https://github.com/HTR-United

Kiessling, B. (2020). A Modular Region and Text Line Layout Analysis System. *17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*.

Kiessling, B. (2022). *Kraken* [source code] http://github.com/mittagessen/kraken

Kiessling, B. (2022). *Kraken* [project website]. http://kraken.re/

Kiessling, B., Stökl Ben Ezra, D., Miller M. (2019) BADAM: A Public Dataset for Baseline Detection in Arabic-script Manuscripts , *HIP@ICDAR* 2019. https://arxiv.org/abs/1907.04041

Kiessling, B.; Tissot, R., Stökl Ben Ezra, D., Stokes P. (2019) eScriptorium: An Open Source Platform for Historical Document Analysis, *OST@ICDAR 2019*. doi: 10.1109/ICDARW.2019.10032

Kiessling, B. (2019) Kraken – A Universal Text Recognizer for the Humanities. *Digital Humanities Book of Abstracts*. doi: 10.34894/Z9G2EX

OCR-D: DFG-Funded Initiative for Optical Character Recognition Development. https://ocr-d.de

OCR/HTR Model Repository. https://www.zenodo.org/communities/ocr_models

SegmOnto. https://github.com/SegmOnto

Stokes, P.A. (2020). RESILIENCE tool: eScriptorium. https://www.resilience-ri.eu/blog/resilience-tool-escriptorium/. Version française : eScriptorium : un outil pour la transcription automatique des documents. https://ephenum.hypotheses.org/1412

Stokes, P., Kiessling, B., Tissot, R., Stökl Ben Ezra, D. (2019) EScripta: A New Digital Platform for the Study of Historical Texts and Writing, *Digital Humanities 2019 Book of Abstracts*. doi: 10.34894/BIXSWX

Stokes, P.A., Kiessling, B., Stökl Ben Ezra, D., Tissot, R. and Gargem, H. (2021). The eScriptorium VRE for Manuscript Cultures. *Ancient Manuscripts and Virtual Research Environments*, ed. Claire Clivaz and Garrick V. Allen. Special issue of *Classics@* 18. https://classics-at.chs.harvard.edu/the-escriptorium-vre-for-manuscript-cultures/

All URLs last verified 21 April 2022.

Notes

1. The source code is released under an MIT licence at https://gitlab.com/scripta/escriptorium/.
2. The source code is released under an Apache 2.0 licence at http://github.com/mittagessen/kraken.

Text and data mining for East Asian sources in classical Chinese

Sturgeon, Donald
djs@dsturgeon.net
Durham University

Brief description

Substantial volumes of primary sources important to the historical written record of China and other East Asian civilizations have been scanned and made available through online databases. Amongst these, the contents of many important sources have been transcribed into textual form, while many more remain available only as images with uncorrected OCR transcriptions. A small but growing number of texts have been semantically annotated, with named entities explicitly marked in the texts and linked to open knowledge bases. Using the Chinese Text Project (https://ctext.org) as a source, this interactive workshop introduces participants to ways of efficiently working

with digitized and annotated historical texts, as well as demonstrating how to improve the state of digitization of such texts in a crowdsourced environment supporting manual correction of OCR, semantic annotation of named entities, and construction and use of a Linked Open Data knowledge graph.

This session will introduce participants to:

1. Basic navigation of this large and moderately complex digital library – e.g. handling of multiple editions, complex metadata etc.
2. Text mining using openly available browser-based tools that use interactive visualizations to allow user-driven exploration of the contents of both this digital library, and arbitrary user-supplied materials in any language.
3. Hands-on introduction to crowdsourced editing to correct errors in textual transcriptions – such as errors introduced through OCR – and principles of versioned textual repositories.
4. Semantic annotation and knowledge base construction. This will introduce the motivation of semantic annotation with concrete examples, and equip participants with the tools to contribute directly to the annotation of classical Chinese sources through crowdsourcing, as well as to the construction of a crowdsourced knowledge graph of data extracted from these same materials.
5. Basic knowledge graph querying and data mining. The knowledge graph introduced supports online querying, the semantics and use of which will be explained in this section.
6. Introduction to querying the knowledge graph with RDF and SPARQL. The knowledge graph introduced closely follows the design principles used in Wikidata, and as such has an RDF representation [1] which can be queried in substantially the same way using SPARQL. This section will provide a brief introduction to this process.

Participants are encouraged to create a free account on ctext.org prior to the workshop by visiting this page: https://ctext.org/account.pl?if=en .

Target audience

Scholars of East Asian history in fields where important source materials are written in classical Chinese (including in particular: China, Japan, Korea, and Vietnam), with interests in any period from around the first millennium BC to 1911 AD. Note that although the source materials used are in classical Chinese, all software used has complete English and Chinese interfaces, and the workshop content should be intelligible to anyone with a minimal degree of Chinese language ability, and/or familiarity with any language written using Sinitic characters (e.g. modern Japanese). Due to the regional importance of classical Chinese historically, sources written in the classical Chinese language remain important in many East Asian historical domains of study.

Motivation

While many researchers working with these materials are likely to have used the Chinese Text Project before – it is accessed by over 30,000 unique users each day – most will not have experience of either the text mining or data mining extensions available, which require a greater investment of effort to meaningfully engage with.

Outline

1. ~20 mins Introduction and overview
2. ~40 mins Interactive text mining using Text Tools (Sturgeon 2018a) and the ctext API (Sturgeon 2021a)
3. ~20 mins Collaborative editing and correcting errors
4. ~40 mins Semantic annotation and knowledge base construction
5. ~30 mins Basic knowledge graph querying and data mining
6. ~30 mins Brief introduction to querying the knowledge graph with RDF and SPARQL

A number of previous workshops run by the same instructor have variously covered many aspects of the material in parts 1 through 5 above, e.g.:

https://dsturgeon.net/aas2021/

https://dsturgeon.net/maraas/

Online written tutorials (created in part for use in previous workshops) exist for much of the content in parts 1, 2, and 3 (available in English, Japanese, and Chinese): https://dsturgeon.net/tutorials/

Part 6 of the tutorial will be entirely new, as RDF serialization of the knowledge graph is a relatively new feature, and previous shorter workshops have lacked sufficient time to cover this aspect.

Instructor

Donald Sturgeon is Assistant Professor of Computer Science at Durham University, and the creator of ctext.org. His research interests include digital libraries, text and data mining, natural language processing of premodern Chinese, and classical Chinese philosophy.

Bibliography

Sturgeon, D. (2018a). Digital Approaches to Text Reuse in the Early Chinese Corpus. *Journal of Chinese Literature and Culture*, **5**(2). Duke University Press: 186–213.

Sturgeon, D. (2018b). Large-scale Optical Character Recognition of Pre-modern Chinese Texts. *International Journal of Buddhist Thought and Culture* doi:10.16893/IJBTC.2018.12.28.2.11.

Sturgeon, D. (2018c). Unsupervised identification of text reuse in early Chinese literature. *Digital Scholarship in the Humanities*, **33**(3): 670–84 doi:10.1093/llc/fqx024.

Sturgeon, D. (2020). Digitizing Premodern Text with the Chinese Text Project. *Journal of Chinese History*, **4**(2). Cambridge University Press: 486–98 doi:10.1017/jch.2020.19.

Sturgeon, D. (2021a). Chinese Text Project: A dynamic digital library of premodern Chinese. *Digital Scholarship in the Humanities*, **36**(Supplement_1): i101–12 doi:10.1093/llc/fqz046.

Sturgeon, D. (2021b). Constructing a crowdsourced linked open knowledge base of Chinese history. *2021 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)*. pp. 1–6 doi:10.23919/PNC53575.2021.9672294.

Notes

1. https://ctext.org/tools/linked-open-data

Git for Humanists: Versioning Research and Code

Tagliaferri, Lisa
lisa.tagliaferri@gmail.com
Rutgers University

This workshop will provide an overview of collaborating on computational humanities projects in Git. A distributed version control system, Git helps both individuals and teams contribute to and maintain software projects, and is very popular within open source. Git can also be useful for version control of text, and we'll explore several ways that humanities researchers can leverage Git effectively.

We'll go over the Git ecosystem, terminology, and how to open a pull request on a team project. Additionally, we'll discuss the various players in the Git space, including GitHub, GitLab, and Bitbucket, and what alternatives there are outside of big tech.

The workshop will also cover an introduction to markdown for those who may be working primarily in text instead of code. We'll also go through the review process and how to add to others' contributions for truly collaborative work.

It is recommended to have a GitHub account and GitHub Desktop installed prior to attending this workshop.

How to Set Up a Web Server for Teaching and Research in the Humanities

Tagliaferri, Lisa
lisa.tagliaferri@gmail.com
Rutgers University

This workshop will go over how to complete an initial Linux server setup for use with the web. We will go over security, firewalls, HTTPS, and high availability. Administering one's own server rather than relying on managed web hosting empowers researchers, teachers, and students by providing them with complete control over their web assets. The resulting setup can be used for webapps, static sites like Jekyll and Hugo, or more robust sites like WordPress, Omeka, Scalar, and Drupal. These will be ready for use with domain names. In addition to providing an entry point to the web, servers can also enable teams of researchers and students to collaborate on programming projects or access shared data.

Panels

Computer Vision for the Study of Printers' Ornaments and Illustrations in European Hand-Press Books

Bahier-Porte, Christelle
christelle.porte@univ-st-etienne.fr
Université Jean Monnet, Saint-Etienne, France

Bergel, Giles
giles.bergel@eng.ox.ac.uk
University of Oxford, United Kingdom

Dutta, Abhishek
adutta@robots.ox.ac.uk
University of Oxford, United Kingdom

Fournel, Thierry
fournel@univ-st-etienne.fr
Université Jean Monnet, Saint-Etienne, France

Thomas, Drew
dt30@st-andrews.ac.uk
University of St Andrews, United Kingdom

Vial-Bonacci, Fabienne
fabienne.vial@univ-st-etienne.fr
Université Jean Monnet, Saint-Etienne, France

Wilkinson, Hazel
h.j.wilkinson@bham.ac.uk
University of Birmingham, United Kingdom; Alan Turing Institute, United Kingdom

Zisserman, Andrew
az@robots.ox.ac.uk
University of Oxford, United Kingdom

Denis, Loïc
loic.denis@univ-st-etienne.fr
Université Jean Monnet, Saint-Etienne, France

Emonet, Rémi
remi.emonet@univ-st-etienne.fr
Université Jean Monnet, Saint-Etienne, France

Habrard, Amaury
amaury.habrard@univ-st-etienne.fr
Université Jean Monnet, Saint-Etienne, France

Ventresque, Vincent
vincent.ventresque@ens-lyon.fr
Université Jean Monnet, Saint-Etienne, France

Gautrais, Thomas
thomas.gautrais@univ-st-etienne.fr
Université Jean Monnet, Saint-Etienne, France

Panel Abstract: Printers' ornaments are the decorative images that appear in printed books to embellish title pages, headings, chapter endings, and any otherwise blank spaces. They were common throughout the hand press period in Europe (c.1470-1830), when they were printed from designs cut into wood or metal blocks, or cast in type-metal (Wilkinson, 2019). They are closely related to woodcut chapbook illustrations, which provided contextual visual stimuli to readers, and, like ornaments, were often designed in such a way that they could be reused in different contexts. Both cuts and type ornaments are like bibliographical fingerprints: their unique features can allow us to identify the printer of a book, even if the printer did not sign their name, or used a pseudonym (Maslen, 2001; Blayney 2021). Identifying a book's printer can help us to date printed material, and better understand the workings of the book trade, and the circulation of texts and ideas. In literary studies, printers' ornaments assist in our enquiries into the original circumstances of a book's production, guiding the decisions of scholarly editors by increasing our understanding of the relationships between authors and the craftspeople who gave their texts material form. Ornaments are also of intrinsic interest as examples of graphic design.

Computer vision and machine learning can help us to build databases of ornaments and cuts, and to derive useful information from them. The four papers in this panel will present different problems and solutions arising from the use of visual AI to investigate ornaments and illustrations. Four projects are represented here, each of which has made some use of the open source VGG Image Search Engine (VISE), and a key aspect of the panel will be discussion of the strengths and limitations of VISE for different kinds of research into ornament and illustration. The papers argue that different approaches are necessary when dealing with large, catch-all databases, compared to smaller, curated datasets that concentrate on a single printer or document type, with two examples of each kind. Paying attention to recent scholarship, several of the papers advance the established potential of ornaments to solve problems of printer identification and fraud detection (May, 2019).

These papers will discuss the strengths and limitations of visual AI *versus* the human expert when it comes to making fine distinctions between images. It is possible to use data visualisation to draw wider conclusions from ornament usage about the activities and habits of printers, which could shape our understanding of the circulation of ideas in early modern Europe. Some of the papers will present arguments for the best use of the data these projects are producing, but the panel will also reflect on the inherent biases and flaws in digitization projects, which can skew our conclusions (Orr, 2021). The panel itself consists of both senior and early career researchers of different nationalities, and includes both men and women.

Paper 1

Title: *Compositor*, Visual AI, and Quantitative Network Analysis: Opportunities and Obstacles

Author: Hazel Wilkinson

Affiliations: Department of English, University of Birmingham, UK; Alan Turing Institute, UK.

Abstract: *Compositor* is a database of over 1 million printers' ornaments and small illustrations from eighteenth-century British books, which was created from *Eighteenth-Century Collections Online* (*ECCO*). Its development was described in Briggs, Gorissen and Wilkinson, 2021. This paper describes how visual AI is being used to further develop *Compositor*, and advances significantly on the 2021 article by investigating new directions in research with the database. *Compositor*is regularly cited by individual scholars as having facilitated their identification of an unknown printer (e.g. Warren *et al*., 2021, Levy, 2021); this paper investigates the possibility of combining VISE's visual search engine with methods from Quantitative Network Analysis (QNA) to use *Compositor*to identify unknown printers on a massive scale (Ahnert *et al.*, 2020). However, first certain qualitative problems with the database must be addressed. The paper builds on Gregg (2020) and Orr (2021) to identify the inherent omissions, errors, and biases that *Compositor* has inherited from *ECCO* and, ultimately, from the *English Short Title Catalogue* (*ESTC*), which complicate using computer vision and QNA to automatically identify printers. The paper will propose strategies for mitigating and acknowledging such gaps and biases that might enable us to unlock the enormous potential of *Compositor* to transform the field of book history.

Paper 2

Title: Using Artificial Intelligence to Identify the Counterfeit Printers of the Protestant Reformation

Author: Drew B. Thomas

Affiliations: University of St Andrews

Abstract: During the 1520s, at the height of the polemical pamphlet campaigns of the Protestant Reformation, one in four of Martin Luther's books was a counterfeit. Printers across the Holy Roman Empire falsely stated on their title pages that their books came from Wittenberg, the home of Luther's movement. Several of these books passed into modern institutions with their true origins still unknown (Künast, 1997, Thomas, 2021). While there are several ways to uncover the printers of counterfeits (by typeface, woodcut or watermark analysis), adopting machine learning and artificial intelligence methods has proven effective. The Ornamento project at University College Dublin is an atlas of the visual geography of the early modern book. Based on forty percent of all known European books printed in the fifteenth and sixteenth centuries (ca.184,000 items), Ornamento has created a record of six million ornamental features, including: musical notation, printers' devices, ornate letters, fleurons, maps, portraits, and illustrations. The results of our efforts allow us to suggest places of publication and printers for a large number of anonymous and counterfeit publications in the period, and to trace how blocks from letters to illustrations have been passed from printer to printer, and in some cases from city to city, region to region. This paper argues that there are clear opportunities to offer clues to assist in unsolved bibliographical mysteries, or to uncover previously hidden networks and associations. For the counterfeit works of Luther, Ornamento allows us to uncover the printers helping spread Europe's first mass media event.

Paper 3

Title: Regions of interest to investigate after learning the use of ornaments by Marc-Michel Rey

Authors: Christelle Bahier-Porte [1], Thierry Fournel [2], Fabienne-Vial Bonacci [1], Loïc Denis [2], Rémi Emonet [2], Amaury Habrard [2], Vincent Ventresque [1], Thomas Gautrais [2]

Affiliations: [1] *IHRIM* - Institut d'Histoire des Représentations et des Idées dans les Modernités, UMR 5317, France

[2] Laboratoire Hubert Curien, UMR 5516, U. Lyon - Université Jean Monnet Saint-Etienne, CNRS, Institut d'Optique Graduate School, France

Abstract: The *Rey Ornament Image investigation* project aims to enable unsupervised novelty/anomaly localization through mapping to help human experts in the authentication of 18th century books (Corsini, 1999, Wilkinson, 2013). In this project, ornaments are not considered in the identification of printers but in the

attribution to a publisher of Enlightenment philosophers named Marc-Michel Rey, at a time when publishing was subject to a censorship regime, and booksellers (such as M.-M. Rey) resorted to anonymity, the use of false addresses, and produced or fell victim to forgeries (Bahier-Porte and Vial-Bonacci, 2020). Ornaments are considered as pieces of evidence when they are single block, and as style marks when composed. Therefore, the aforementioned machine learning task is based on a limited dataset of a few hundred Rey ornaments.

A database was designed with the collected Rey ornaments in order to cross-reference potential anomalies with other investigable clues. By querying the database with a current ornament, similar ornaments can be extracted online using VISE (Visual Image Search Engine) (Bergel *et al*., 2020). From that point, our paper argues that we can suggest addressing novelty localization by learning normal local image variations or how to reconstruct normal images, both from a small set of retrieved images. Different types of maps can be derived as heatmap (Li *et al*., 2021), difference map (Baur *et al.*, 2021) or attention map (Venkataramanan *et al.*, 2020), enabling a more or less well-resolved visualization of novelties in the current query. Rey's emblematic mark is analyzed to illustrate our argument.

Paper 4

Title: Visual Analysis of Chapbooks Printed in Scotland

Authors: Abhishek Dutta, Giles Bergel, Andrew Zisserman

Affiliations: Department of Engineering Science, University of Oxford

Abstract: Chapbooks were short and cheaply produced reading material (e.g. songs, poems, stories, games, riddles, religious writings) that were available from the end of 17th century to the late 19th century (Fox, 2013). These chapbooks were designed to appeal to a wide readership and they often contained illustrations which provided contextual visual stimuli to the readers (Ross Roy, 1974). These illustrations often appeared in the title pages (Beavan, 2019). This paper describes the visual analysis of such chapbook illustrations. We automatically extract all the illustrations contained in 3000 chapbooks printed in Scotland and create a visual search engine to visually search this collection using full or part-illustrations as search query. We also group these illustrations based on their visual content, and provide keyword-based search of metadata associated with each publication. The visual search, grouping of illustrations based on visual content, and metadata search features enable researchers to forensically analyse the chapbooks dataset and to discover unnoticed relationships between its elements. We release all annotations and software tools described in this paper to enable reproduction of the results presented and to allow extension of the methodology described to datasets of a similar nature (Dutta *et al.*, 2021). We also show how these tools are being used for analysis of other chapbook datasets (e.g. the Spanish Chapbooks) of similar nature.

Bibliography

Ahnert, R., Ahnert, S., Coleman, C. N., Weingart, S. B. (2020), *The Network Turn: Changing Perspectives in the Humanities*. Cambridge.

Bahier-Porte, C. and F. Vial-Bonacci (2020), «Le commerce de la librairie à la lumière de la correspondance – Marc Michel Rey, Pierre rousseau, Charles Weissenbruch, F. Tilkin *(dir.)* » *Trois siècles d'histoire du livre et de la pensée à travers le Fonds Weissenbruch - Du Journal encyclopédique aux humanités numériques,* Bruxelles, Archives générales du Royaume, coll. *Studia*, n°166, p. 205-222.

Baur, C., Wiestler, B., Albarqouni, S., and Navab, N. (2018), «Deep autoencoding models for unsupervised anomaly segmentation in brain MR images.» In *International MICCAI Brain Lesion Workshop*, pp. 161-169. Springer, Cham.

Beavan, I. (2019), 'Lines of defence: thoughts on Scottish chapbook title-page woodcuts and their functions'. *Publishing History* 81, pp. 41–5.

Bergel, G., Franklin, A., Heaney, M., Arandjelovic, R., Zisserman, A. and Funke, D. (2013), «Content-based image recognition on printed broadside ballads: The Bodleian Libraries' ImageMatch Tool» Proceedings of the IFLA World Library and Information Congress.

Corsini, S. (1999), «La preuve par les fleurons?: analyse comparée du matériel ornemental des imprimeurs suisses romands: 1775-1785» Ferney-Voltaire, Centre international d'étude du XVIII e siècle.

Blayney, P. W. M. (2021), 'The Flowers in the Muses Garland'. *The Library* 22.3, pp. 316–43.

Dutta, A., Arandjelović, R., Zisserman, A. (2021) 'VGG Image Search Engine'. Retrieved 8/12/21 from https://www.robots.ox.ac.uk/~vgg/software/vise/

Fox, A. (2013), '"Little Story Books" and "Small Pamphlets" in Edinburgh, 1680– 1760: The Making of the Scottish Chapbook', *Scottish Historical Review* 92, pp. 207–230.

Gregg, S. H. (2020), *Old Books and Digital Publishing: Eighteenth-Century Collections Online*. Cambridge.

Künast, H. (1997), *Getruckt zu Augsburg: Buchdruck und Buchhandel in Augsburg zwischen1468 und 1555*. Niemeyer, pp. 167–168.

Levy, D. (2021), 'Who printed Piquet for Francis Cogan?'
https://edmondhoyle.blogspot.com/2021/01/who-printed-piquet-for-francis-cogan.html

Li, C. L., Sohn, K., Yoon, J., & Pfister, T. (2021), «CutPaste: Self-Supervised Learning for Anomaly Detection and Localization» In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9664-9674.

Maslen, K. I. D. (2001), *Samuel Richardson of London, Printer*. University of Otago Press.

May, J. E. (2019), 'False and Incomplete Imprints in Swift's Dublin, 1710–35', in Bischof, Juhas & Real (eds), *Reading Swift: Papers from the Seventh Münster Symposium on Jonathan Swift*. Brill.

Orr, L. (2021), 'From Methods to Conclusions: The Limits of the Knowable in Digital Book History'. *Eighteenth Century Studies* 54.4, pp. 785–801.

Ross Roy, G. (1974), 'Some notes on Scottish chapbooks', *Scottish Literary Journal* 1, pp. 50–60.

Thomas, D. (2021), *The Industry of Evangelism: Printing for the Reformation in Martin Luther's Wittenberg*. Brill.

Venkataramanan, S., Peng, K. C., Singh, R. V., & Mahalanobis, A. (2020), «Attention guided anomaly localization in images» In *European Conference on Computer Vision*, pp. 485-503. Springer, Cham.

Warren, C. N., Wiscomb, A., Williams, P., Lemley, S. V., G'Sell, M. (2021), 'Canst Thou Draw Out *Leviathan*with Computational Bibliography? New Angles on Printing Thomas Hobbes' "Ornaments" Edition'. *Eighteenth-Century Studies* 54.4, pp. 827–59.

Wilkinson, H. (2013), «Printers' Flowers as Evidence in the Identification of Unknown Printers: Two Examples from 1715» *The Library*, 7th ser., 14, p. 70-79.

Wilkinson, H (2019), 'Printer's Ornaments and Flowers ', in A Smyth & D Duncan (eds), *Book Parts.* Oxford University Press.

Wilkinson, H, Briggs, J & Gorissen, D. (2021), 'Computer vision and the creation of a database of printers' ornaments', *Digital Humanities Quarterly*, vol. 15, no. 1.

The European Open Science Cloud (EOSC) and Its Implications for the Digital Humanities and Social Sciences

Barbot, Laure
laure.barbot@dariah.eu
DARIAH-EU

Gray, Edward
edward.gray@dariah.eu
Huma-Num CNRS

Fischer, Frank
fr.fischer@fu-berlin.de
Freie Universität Berlin; DARIAH-EU

Broeder, Daan
daan.broeder@di.huc.knaw.nl
CLARIN ERIC

Ďurčo, Matej
matej.durco@oeaw.ac.at
Austrian Centre for Digital Humanities and Cultural Heritage

Kleemola, Mari
Mari.Kleemola@staff.uta.fi
Finnish Social Science Data Archive at Tampere University

Thiel, Carsten
carsten.thiel@cessda.eu
CESSDA ERIC

König, Alexander
alex@clarin.eu
CLARIN ERIC

Introduction

(Frank Fischer)

The European Open Science Cloud (EOSC), officially launched in November 2018, is an initiative of the European Commission aiming to develop an infrastructure that makes access to scientific data easier and more efficient and is geared towards open science. The planned infrastructure bundles services of different providers and is following a system-of-systems approach. Under the Horizon 2020 framework programme, five cluster projects from the sciences are being funded to connect research communities of different fields to the EOSC. One of these projects is SSHOC – the Social Sciences and Humanities Open Cloud.

The five European Research Infrastructure Consortia CESSDA, CLARIN, DARIAH, ESS and SHARE as well as 20 other European partner institutions have come together to connect their communities to the EOSC via SSHOC. In

this panel, we will discuss its emerging role for the Digital Humanities and related fields and the different facets needed to design, coordinate and operate this domain-specific infrastructure and cloud. A panel of speakers will discuss the following topics:

The SSH Open Marketplace – Contextualised practical knowledge for day-to-day research in the Digital Humanities

(Laure Barbot, Edward Gray, Klaus Illmayer, Alexander König)

Tool directories and discovery portals have a long history in the Digital Humanities. There is a clear demand for them in the community but their sustainability beyond project fundings and their collective maintenance is challenging (Dombrowski, 2019; Edmond, 2016). In the enabling EOSC environment, in which research and data infrastructures find a suitable framework to provide research services, DARIAH, CLARIN and CESSDA developed the SSH Open Marketplace, accessible at https://marketplace.sshopencloud.eu/, a new take on discovery platforms for tools, services, training materials, datasets and digital workflows (Kálmán et al., 2019). While crowd-based maintenance of content is one of the few ways to ensure the quality of a "social marketplace for services" (Blanke, 2011), the SSH Open Marketplace relies on a set of rewards and incentives and a small, but funded, editorial team in order to avoid pure volunteer work of its contributors.

The Sharing of Services and Data Between SSH Infrastructures

(Daan Broeder)

By sharing infrastructure resources we gain considerable advantages:

- (1) scaling advantage: increasing the potential use and impact of (expensive) datasets and maximizing the investment needed for developing services and
- (2) offering SSH researchers new choices and opportunities with regards to data creation and analysis.

This concerns not only newly created services and datasets, e.g. the SSH Open Marketplace, but also existing resources which were prepared to serve a wider audience. Within the SSHOC project, a couple of services originally developed for the CLARIN community were generalised to address additional user scenarios and provide new functionalities. Recommendations and tools for creating and managing SSH vocabularies were aligned through an SSH Vocabulary Commons initiative. SSHOC also provided an opportunity for new and emerging communities to discover and investigate what and how already established communities have arranged their infrastructure and what is useful for them.

Data Repository/Service Certification

(Mari Kleemola)

FAIR data (Findable, Accessible, Interoperable, Reusable) is essential for the success of the EOSC, but the demand for high quality and FAIR data and metadata is global. However, the FAIR principles do not take into account the inevitable changes in the data environment and in the needs of data users. Any digital system that accepts, stores and provides access to data can be termed a 'digital repository' in the broadest sense. Only an organisation that meets clear criteria, including the provision of active long-term preservation measures for a designated community, can be termed a “trustworthy digital repository”. Trustworthy digital repositories play a key role in enabling data to become and remain FAIR over time within all disciplines, including SSH. Reaching FAIRness and trustworthiness is a journey that requires expertise, investment and cooperation. Many actors and multiple tiers of data storage, protection and preservation are needed to address the long tail of at-risk data and metadata. Setting standards and pass/fail criteria is not sufficient; compliance with trustworthiness standards is best supported by community efforts such as repository certification through CoreTrustSeal, existing coordination networks for digital repositories in disciplinary contexts and other global, regional and national networks.

Technology and Services

(Carsten Thiel)

Research Infrastructures are increasingly digital, providing virtual access to resources at physical places like libraries and archives. Developing the technology that makes this possible is a fundamental part of building modern infrastructures and enabling interdisciplinary research requires interoperability on the technical level. But delivering added value to the researcher requires more than bare technology. In collaboration across disciplines, both within the SSHOC project and the wider EOSC landscape, the focus has shifted towards the service layer, combining tools with training and support into a complete package that addresses end user needs. As domain infrastructures, CESSDA, CLARIN and DARIAH are uniquely positioned

between researchers, resources and tools to process them in order to sustainably ensure open research.

EOSC FUTURE – Connecting Scientific Projects to EOSC Core Services

(Matej Ďurčo)

To connect thematic services to the EOSC basic agnostic components are needed. The EOSC Architecture provides the so-called EOSC-Core layer offering fundamental building blocks such as an AAI (Authentication and Authorization Infrastructure), Helpdesk and Monitoring. By easily integrating these core services, science projects benefit from an interoperable and more stable infrastructural environment. The domain-oriented cluster projects are at the crossroads between agnostic service providers and research communities to overcome the "last mile challenge" (Koureas, 2016) and build mid-level tools and platforms that serve the needs of individual research communities (Lamanna, 2021).

Modelling and Operationalizing Concepts in Computational Literary Studies

Brandes, Phillip
phillip.brandes@uni-jena.de
University of Jena, Germany

Dennerlein, Katrin
katrin.dennerlein@uni-wuerzburg.de
University of Würzburg, Germany

Jacke, Janina
janina.jacke@uni-goettingen.de
University of Göttingen, Germany

Marshall, Sophie
sophie.marshall@uni-jena.de
University of Jena, Germany

Pielström, Steffen
pielstroem@biozentrum.uni-wuerzburg.de
University of Würzburg, Germany

Schneider, Felix
felix.schneider@uni-jena.de
University of Jena, Germany

Introduction (Steffen Pielström)

In computational literary studies (CLS), research questions about literature are approached with computational, often quantitative research methods. This approach often requires researchers to model and operationalize concepts from literary studies in a way that renders them accessible to quantitative methodology. Research in literary studies traditionally follows a qualitative paradigm. Concepts are often defined more vaguely than, for example, in the natural sciences, and often subject to continuing controversies. This practice results in a rich repertoire of terms that allows for the description of a very broad spectrum of phenomena by readers of literature. In the natural sciences on the contrary, definitions are often from the very beginning measured by their possibility to be operationalized for mathematical treatment, effectively limiting choice of concepts that can be objects of research.

As a consequence, CLS can not always easily represent the entire spectrum of concepts and research questions present in literary studies. Sometimes methodological opportunities suggest different focuses. For example, computational studies on writing style still often focus on authorship, a text attribute that can be determined for a majority of literary works without much controversy. Beyond authorship analysis, literary genre is a phenomenon often looked at, and whereas early studies tended to rely on a very simplifying, categorical concept of genre, recent works try to establish new and more convincing genre concepts from a CLS point of view (see Henny-Krahmer et al. 2018, Calvo-Tello 2020).

Another example can be found in a current project on the complexity of literary texts (Pielström et al. in preparation). Though a concept of complexity is well established in the natural sciences, it proved to be surprisingly complicated to model and operationalize literary scholar' concepts on the complexity of a text concisely in a way fit for mathematical treatment. Even after the extreme narrowing of the project's focus on lexical complexity, there remains a number of dimensions to the problem, and each possible operationalization is weighting these dimensions differently. Ultimately, each measuring approach can only represent one point of view on the multiple dimensions of lexical complexity and remains thereby imperfect (see also Jarvis 2013).

As these examples show, modelling concepts from literary studies for computational, quantitative methods can mean walking a thin line between precision and relevance: On one hand, concepts often can only be operationalized

by reducing and simplifying them, controversial concepts can sometimes only be treated by relying on pragmatic working definitions. On the other hand, doing work on the basis of simplifying definitions can lead to results that are not considered relevant by other literary scholars any more. With this panel, we aim to invite a broader discussion on how CLS researchers can or should deal with this issue and how it potentially influences the relationship between CLS and other parts of the literary studies community. As a stimulus for this discussion, we present examples from different CLS projects and their solutions for modelling literary studies concepts.

Modelling concepts of literary science - a typology (Phillip Brandes and Sophie Marshall)

The complexity of literary texts continues to pose regular challenges to computational literary studies. An essential part of computational work with literary texts consists of modelling and operationalising phenomena of literary texts. Phenomenon is understood broadly here and includes, for example, stylistic properties of texts such as certain stylistic devices, abstract, narratological concepts such as the narrative instance or even everyday world themes that are negotiated in texts such as emotions.

With this contribution we propose a preliminary approximative typology of what can (already) be modelled and what cannot (yet). For this purpose, we make a division into, on the one hand, 1) "simple" text(phenomena) and 2) "complex" text(phenomena) and, on the other hand, a) "simple" theoretical concepts and b) "complex" theoretical concepts. This results in four categories (1a, 1b, 2a and 2b), of which only category 1a seems to be modelable without major difficulties. An example would be the stylistic device of anaphora - "Show all at least 3-word strings beginning with identical letters" is implementable with comparatively little effort.

Intuitively, one would thus say problems in modelling increase with the complexity and ambiguity of what is to be modeled. Categories 2a) and 2b) are of interest for the planned contribution. For the latter, we would need to ask to what extent we are approaching the possibility of modelling complex or ambiguous concepts.

Stylistic devices and their function for the representation of 'Asia' in Middle High German literature (Felix Schneider and Phillip Brandes)

As an example of category 2a) we will use the stylistic figures parallelism and chiasmus (Fauser 1994, Ostrowocz 2003). We operationalise the latter as inversions of part-of-speech tags according to an A B B' A' scheme (which includes antimetabole, an inverted repetition of lemmas, as a special case of chiasmus), the former according to the A B A' B' scheme, and detect them using a machine learning classifier pre-trained on German chiasmi (Schneider et al. 2021). The particular challenge here is to distinguish between random part-of-speech tag inversions and true stylistic devices. The metaphor serves as an example of category 2b) that is more complex (Goldmann, 2019) than chiasmi and parallelisms. Our approach, therefore, is to first model only certain types of metaphors. The results of this approach are so far preliminary and serve only as a basis for further discussion.

We finally examine the stylistic devices modelled in this way in a corpus of Middle High German texts that can only approximate the diversity of Asia in their engagement with non-European spaces (e.g., "Willehalm" by Wolfram von Eschenbach, "Herzog Ernst", and "König Rother"; of which the thematization of Asia, especially India and parts of the Near East, has already been studied using conventional methods: Prager 2014, Schmitz 2018 and Sivri 2016). The article thus stimulates a (heuristic) typology for modelling concepts of literary studies, tests their implementation, and sheds light on the diversity of Asia from the perspective of Medieval German literature.

Modelling emotions in drama - literary studies and digital methods (Katrin Dennerlein)

In the "Emotions in Drama"-project we are investigating how the emotions of characters were expressed in 17th to early 19th century German drama. We want to answer questions like: how do subgenres differ in terms of distributions and development of expressed emotions? Can diachronic developments be modeled based on emotions? Can these methods help to identify groups of texts in unknown materials that were considered as being related by contemporaries? The basis for answering these questions is a method that identifies emotions in character speech. To establish such a method, a neural network classifier is trained to recognize emotions in annotated texts (Schmidt/Dennerlein/Wolff 2021). We selected 13 emotions particularly important for the drama of that period (Schings 1980, Meier 1993, Schulz 1988; Schonlau 2017; Dennerlein/Schmidt/Wolff 2022) and annotated them in terms of source, target and polarity (Kim/Klinger 2019). So far, we collected about 20,000 instances of emotions in 17 dramas from our study period. After using these annotations to train Gbert we are currently classifying sentences of dramatic texts which the trained model hasn't seen before. In this step, differentiations in literary studies such as the

distinction between presentation/thematisation (Schonlau 2017) and the diversification of various attribution and classification steps must still be left aside (Gius/Jacke 2017, Schonlau 2017, 34-38). As soon as it is clear with which training data and transformer models neural networks best learn to "understand" historical drama language, complex and literature-specific phenomena such as metaphors, lies, irony, pathos or comedy are to be modelled.

Modelling interpretation-dependent concepts - the example of the unreliable narrator (Janina Jacke)

From a CLS perspective, heavily formalized concepts based on descriptive text features are most attractive for computational analysis. From a literary studies perspective, however, such concepts often represent only heuristic, auxiliary constructions that assist text interpretation (see Kindt/Müller 2003). But there is a value in modelling more complex phenomena. The attempt to model complex concepts requires theoretical analysis investigating which aspects are strictly descriptive and which rely on contextual knowledge and preceding conclusions (Gius/Jacke 2017). This analysis is in itself valuable for literary studies, since it renders method and theory more transparent. In the best case, this allows for a computational treatment of the descriptive aspects. Even research on more complex topics can thereby profit from the formalist perspective of information science and use input from computational analysis without simplification.

The phenomenon of the unreliable narrator provides an excellent example, since this concept, though interpretation-dependent, does have aspects suitable for computational treatment. In the CAUTION project the concept is treated by discriminating between different subtypes (Jacke 2020). For some of these subtypes, like fact-based unreliability, language-level indicators can be identified that are, at least partially, detectable with computational methods. On this basis quantitative surveys on larger corpora are possible, and their results can be included in interpretation.

As a second step, inference engines are used to compare contrasting interpretation hypotheses for selected texts in terms of consistency, coherence and scope. Adapting computational methods thereby supports the genesis of literary theory and text research without reducing concepts to their constituent formal components.

Bibliography

Calvo Tello, J. (2020). What is a Genre? A Graph Unified Model of Categories, Texts, and Features. Ottawa: *ADHO* https://hcommons.org/deposits/item/hc:31713/ (accessed 12 October 2020).

Dennerlein, K, Schmidt, T. and Wolff, C. (2022): Emotionen im kulturellen Gedächtnis bewahren., in: *Book of Abstracts, DHd2022*.

Fauser, M. (1994): Chiasmus, in: *Historisches Wörterbuch der Rhetorik*, Bd. 2, c. 171–173.

Gius, E. and Jacke, J. (2017): The Hermeneutic Profit of Annotation. On Preventing and Fostering Disagreement in Literary Analysis. In: I *nternational Journal of Humanities and Arts Computing* 11: 2, 233 –254.

Goldmann, L. (2019): *Phänomen und Begriff der Metapher. Vorschlag zur Systematisierung der Theoriegeschichte*. De Gryuter.

Haferland, H. (2019): Erzähler, Fiktion, Fokalisierung. Drei Reizthemen der Historischen Narratologie. In: *BME* 2, S. 11–147.

Henny-Krahmer, U., Betz, K., Schlör, D. a nd Hotho, A. (2018). Alternative Gattungstheorien. Das Prototypenmodell am Beispiel hispanoamerikanischer Romane. *DHd2018*. Universität zu Köln, March 1, 2018.

Jacke, J. (2020): *Systematik unzuverlässigen Erzählens. Analytische Aufarbeitung und Explikation einer problematischen Kategorie*. Berlin/Boston: de Gruyter.

Jarvis, S. (2013): Chapter 1. Defining and measuring lexical diversity. In S. Jarvis & M. Daller (Hrsg.): *Studies in Bilingualism* 47: 13-44. John Benjamins Publishing Company. https://doi.org/10.1075/sibil.47.03ch1

Kim, E. and Klinger, R. (2019): A survey on sentiment and emotion analysis for computational literary studies., in: *Zeitschrift für digitale Geisteswissenschaften*.

Kindt, T. and Müller, H.- H. (2003): Wieviel Interpretation enthalten Beschreibungen? Überlegungen zu einer umstrittenen Unterscheidung am Beispiel der Narratologie. In: Jannidis, F., Lauer, G., Martínez, M. and Winko, S. Eds. (2003.): *Regeln der Bedeutung*. Berlin/New York: de Gruyter, 286–304.

Meier, A. (1993). *Dramaturgie der Bewunderung: Untersuchungen zur politisch-klassizistischen Tragödie des 18. Jahrhunderts* (Vol. 23). Frankfurt am Main: Vittorio Klostermann.

Martínez, M. and Scheffel, M. (2016): *Einführung in die Erzähltheorie*, München: C.H. Beck.

Ostrowocz,

Philipp (2003): Parallelismus, in: *Historisches Wörterbuch der Rhetorik*, Bd. 6, c. 546–552.

Pielström, S., Hodošček, B., Calvo Tello, J., Henny-Krahmer, U., Jannidis, F., Schöch, C., Du, K., Uesaka, A. and Tabata, T. (in preparation): Measuring Lexical Diversity of Literary Texts.

Prager, D.N. (2014): *Orienting the self. The German literary encounter with the Eastern other*. Camden House.

Schings, H.- J. (1980). *Der mitleidigste Mensch ist der beste Mensch: Poetik des Mitleids von Lessing bis Büchner.* Edition Beck. München: Beck.

Sivri, Y. (2016): *Mitteldeutsche Orientliteratur des 12. und 13. Jahrhunderts. 'Graf Rudolf' und 'Herzog Ernst'. Ein Beitrag zu interkulturellen Auseinandersetzungen im Hochmittelalter*. Peter Lang.

Schleiermacher, F. D. E. (1997): Hermeneutik und Kritik mit besonderer Beziehung auf das Neue Testament, In:Frank, W. (ed.): *Hermeneutik und Kritik*. Frankfurt am Main: suhrkamp.

Schmitz, F. (2018): *Der Orient in Diskursen des Mittelalters und im "Willehalm" Wolframs von Eschenbach*. Peter Lang.

Schneider, F., Brandes, P., Barz, B., Denzler, J. and Marshall, S. (2021): Data-Driven Detection of General Chiasmi Using Lexical and Semantic Features. *ACL 2021*. P. 96–100.

Schonlau, A. (2017). *Emotionen im Dramentext: eine methodische Grundlegung mit exemplarischer Analyse zu Neid und Intrige 1750-1800* (Vol. 25). Berlin: Walter de Gruyter GmbH & Co KG.

Schulz, G. M. (1988). *Tugend, Tod und Gewalt. Das Trauerspiel der Aufklärung und die Dramaturgie des Pathetischen und des Erhabenen.* Tübingen: Niemeyer.

Schwarz-Friesel, M. (2007). *Sprache und Emotion*. Tübingen: Francke.

The Politics of Digital Humanities Infrastructure and Sustainability

Burkert, Mattie
mburkert@uoregon.edu
University of Oregon, United States of America

Moore, Shawn
shawn.moore@fsw.edu
Florida Southwestern State College, United States of America

Gil, Alex
agil@columbia.edu
Columbia University, United States of America

Liebe, Lauren
leliebe@tamu.edu
Texas A&M University, United States of America

Nicosia, Marissa
mon4@psu.edu
The Pennsylvania State University – Abington College, United States of America

Otis, Jessica
jotis2@gmu.edu
George Mason University, United States of America

Wikle, Olivia
omwikle@uidaho.edu
University of Idaho, United States of America

Williamson, Evan Peter
ewilliamson@uidaho.edu
University of Idaho, United States of America

Becker, Devin
dbecker@uidaho.edu
University of Idaho, United States of America

Session Overview

In November 2021, a heated discussion broke out on social media about Johanna Drucker's article on "Sustainability and Complexity," which detailed her experiences with digital project migration and obsolescence. While many researchers applauded Drucker for drawing attention to the sustainability problem in DH, information professionals pointed out that they have long recognized and worked to address the challenges of hosting, maintenance, and preservation Drucker identifies. The fact that Drucker's piece cited little of this work reflects a pernicious tendency among humanities scholars to disregard the expertise of librarians, designers, and developers (Posner 2013; @ThatAndromeda). Our panel addresses the widespread degradation of DH projects as a crisis, not of inattention or negligence, but of misalignment between the technical, institutional, and social infrastructures of DH work.

Challenges of project endurance and afterlife management are well documented in DH (for example, see Barats, Schafer, and Fickers 2020). In the wake of high-profile collapses like Project Bamboo (Dombrowski 2014, Almas 2017), funders including ACLS and NEH have offered grants for recovery, extension, and preservation of existing work. In addition, projects like the Socio-Technical Sustainability Roadmap (Visual Media Workshop at the University of Pittsburgh) and the Endings Project (Humanities Computing Media Centre at the University of

Victoria) have emerged to educate teams about the reality of impermanence and to assist with the sunsetting process (see also Otis 2021). Recognizing the labor and environmental costs of maintenance, a growing number of DH thinkers and institutions are rejecting the imperative to keep projects alive forever, embracing ephemerality and a minimalist ethos (Lincoln 2020; Sayers 2016; Gil 2015).

Drawing on insights from critical infrastructure studies, our panel begins from the premise that technical infrastructures are inseparable from social and political concerns (Thylstrup 2018; Deger-Pawlicka 2021). Attempting to address sustainability as a technical challenge without attention to its sociopolitical dimensions leads to tribalism, retrenchment, and re-inscription of existing power imbalances and resource inequities. This panel will therefore explore the sustainability problem in DH as a *matter of concern* (Latour 2005), a complex object constructed by multiple intersecting and competing agencies.

Our panel is led by two faculty project directors and brings together academic researchers, teachers, librarians, information technologists, software designers, engineers, and project managers with a range of experience as DH makers and maintainers. By highlighting the perspectives of different kinds of workers within and adjacent to different kinds of institutions, we hope to identify specific frictions between the material and technical conditions of our work and the sociocultural realities in which it takes place. Together, we will take a fresh look at well-known sustainability issues, reframing them through the lens of critical infrastructure and critical university studies in order to imagine a new set of interconnected disciplinary, institutional, and technical infrastructures that would enable DH scholarship to thrive.

Presentations

Alex Gil, *The Post-Infrastructures of Our Trash Futures*

In this paper, I propose that minimal computing operates in a future and past that repurposes that which has been discarded, trashed, ignored. The dream of a shared and reliable humanities infrastructure has been a mainstay of digital humanities for decades: projects like Humanities Commons, DARIAH, Omeka and others, are results of the pursuit of our collective scholarly independence from corporate solutions and control of our own work. In this paper, I argue for a Plan B where all of our infrastructural dreams fall by the wayside and we are left in the position of scavengers of corporate tech infrastructure and open standards without sacrificing the mission of the humanities —to steward and interpret human culture.

In this talk, I will also directly address the social and political challenge posed by faculty project directors who fail to understand, or understand too late, the real cost of invisible labor, and their often unrealistic expectations around the stewardship of their projects—made vanity by their own failure in transforming tenure & promotion guidelines and other material reward mechanisms in their departments. Within the context of shifting blame to neoliberalism, I will argue for the vacuity of theory absent critical infrastructural practice tied to the means of production of humanistic knowledge. Plan B, I will argue, makes all the more sense within the grim prospect that faculty project directors will continue to behave as they have in the past few decades.

Lauren Liebe, *Flexibility as Sustainability in Digital Humanities Projects*

Johanna Drucker's recent article "Sustainability and Complexity: Knowledge and Authority in the Digital Humanities" raises important questions about knowledge specialization within digital humanities projects. While Drucker asserts that the digital humanities must incorporate humanistic methodologies, not just humanistic content, the reverse is also true: humanist digital projects must also make use of the full potential of their digital expressions. To do this successfully requires that both the humanities scholars and the technical experts possess nuanced understandings of both the project's content and the technology upon which it relies.

Such an understanding allows digital humanities projects to explore the nuances of their work as both digital *and* humanist. One of these nuances, as Drucker points out, is that "we need to think of the work of digital humanities as radically incomplete, always ongoing" (93). While physical media creates a sense of stability, the digital always operates with a level of ephemerality. Even archival projects like the Internet Archive capture only snapshots, not projects in their entirety. Key to this notion of digital ephemerality is the need to embrace flexible technology, particularly in the creation and storage of data and metadata, to allow for a wider user base and interoperability with other projects. In this presentation, I approach this problem from the perspective of my work as the project manager for the Advanced Research Consortium, an aggregator of data from digital humanities, proprietary scholarly resources, and library databases, to discuss how flexible data management enriches digital humanities projects.

Marissa Nicosia, *Secretary Hand, Digital Interface: Sustainable Collaborative Research with Undergraduate Students*

This paper has been withdrawn.

Jessica Otis, *Cui Bono? Costs, Benefits, and Priorities in Digital Sustainability*

Digital humanists are good at making do: cobbling together projects with "free" resources, volunteer labor, time-limited funding, and access to university infrastructures. But while we are generally more aware of economic realities than other humanists—recognizing concepts such as overhead costs and fringe benefit rates—there is still a widespread misunderstanding about the real costs of digital scholarship, especially sustainability costs. In part this is because digital humanists from book-based disciplines have been conditioned to expect scholarly immortality. Books are distributed into university libraries with low marginal costs and survive for decades or centuries. Eventually, librarians must decide if books are valuable enough or being used enough to justify their ongoing preservation costs, but there is no additional work required from the author. Yet unlike books, digital projects have significant sustainability costs that escalate over time and require regular reassessment over how, and how long, to keep them online.

This paper argues digital humanists must learn to understand the real costs of digital sustainability and assess the benefits of keeping projects online. It will frankly discuss RRCHNM's backlist of digital projects, and the socioeconomic and political foundations of our analyses of costs and benefits for sustaining those projects. Crucially, it will filter that discussion through the lens of potential institutional priorities, from avoiding hacking to benefiting current students to advancing certain types of scholarship to generating prestige. When infrastructural resources are finite, we must learn to consider costs, benefits, and priorities when deciding how to employ them for digital sustainability.

Olivia Wikle, Evan Peter Williamson, and Devin Becker, *Using Static Web Methodology as a Sustainable Approach to Digital Humanities Projects*

The web platforms adopted for digital humanities projects come with significant short and long term costs. In the realities of academic funding, this often results in huge sums sunk into outsourced development, contract work, and 3rd party subscriptions, reflecting an economic model that prioritizes purchasing systems over internal development of people and capacity. As DH practitioners, the time (or money paid to contractors) we must invest in managing servers, maintaining platform updates, and learning idiosyncratic administrative systems ultimately limits our ability to create and sustain unique, innovative projects. In response, librarians and DH practitioners are reexamining DH platforms through a minimal computing lens, pursuing new project-development methods that minimize digital infrastructure as a means to maximize investment in people, growing agency, agility, and long term sustainability in both the organization and digital outputs. Eager to explore this potential, faculty librarians at University of Idaho have been developing digital collections, scholarship projects, and instructional content using static web tools for more than five years, beginning with the digital collections template CollectionBuilder and expanding to include projects such as oral history exhibits, deep maps, and digital editions. This development approach, which we call Lib-Static, seeks to increase the return on learning new technical skills that all digital projects require, while also establishing technical solutions and social workflows that more closely match the structure of academic work cycles and DH project needs. In particular, the static web approach encourages the creation of preservation-ready project data, enables periods of iterative development, and capitalizes on the low-cost/low-maintenance characteristics of statically-generated sites to optimize limited economic resources and personnel time. This presentation will introduce the Lib-Static development methodology as a provocation to rethink DH infrastructure choices, asking how our frameworks can build internal skills, collaboration, and empowerment to generate more sustainable digital projects.

Bibliography

@ThatAndromeda. (2021). *Twitter*. 10 Nov 2021, https://web.archive.org/web/20211122202905/https://threadreaderapp.com/thread/1458445616409939971.html (accessed 1 Dec 2021).

Almas, Bridget. (2017). "Perseids: Experimenting with Infrastructure for Creating and Sharing Research Data in the Digital Humanities." *Data Science Journal*, 16: 19, DOI: https://doi.org/10.5334/dsj-2017-019.

Barats, Christine, Valerie Schafer, and Andreas Fickers. (2021). "Fading Away… The challenge of sustainability in digital studies", *DHQ* 14.3, http://www.digitalhumanities.org/dhq/vol/14/3/000484/000484.html (accessed November 22, 2021).

Deger-Pawlicka, Urszula. (2021). "Infrastructuring digital humanities: On relational infrastructure and global reconfiguration of the field. *Digital Scholarship in the Humanities* fqab086: https://doi.org/10.1093/llc/fqab086 (accessed 18 Nov 2021).

Dombrowski, Quinn. (2014). "What Ever Happened to Project Bamboo?" *Literary and Linguistic Computing* 29.3: 326–339 https://doi.org/10.1093/llc/fqu026 (accessed 22 Nov 2021).

Drucker, Johanna. (2021). "Sustainability and Complexity: Knowledge and Authority in the Digital Humanities." *Digital Scholarship in the Humanities*, 36 supp_2: ii86–ii94, https://doi.org/10.1093/llc/fqab025 (accessed 20 Nov 2021).

Gil, Alex. (2015). "The User, the Learner and the Machines We Make." *Minimal Computin,*. http://go-dh.github.io/mincomp/thoughts/2015/05/21/user-vs-learner (accessed 19 Nov 2021).

Humanities Computing Media Centre at the University of Victoria. (2021). *The Endings Project*. https://endings.uvic.ca (accessed 22 Nov 2021).

Latour, Bruno. (2005). *Reassembling the Social: An Introduction to Actor-Network Theory*. Oxford University Press.

Lincoln, Matthew. (2020). "From Supercomputer to Static Site: Boiling Down Big Research Data for Preservation and Usability." code4lib 2020, Pittsburgh, PA, https://kilthub.cmu.edu/articles/presentation/From_Supercomputer_to_Static_Site_Boiling_Down_Big_Research_Data_for_Preservation_and_Usability/18280082 (accessed 2 April 2022).

Otis, Jessica. (2021). "Managing the Digital Backlist: Sustaining, Preserving, and Deleting Old Projects." Digital Humanities Summer Institute Colloquium (virtual). June 15, 2021, https://dhsi.org/timetable/event/institute-lecture-jessica-otis (accessed 15 Nov 2021).

Posner, Miriam. (2013). "No Half Measures: Overcoming Common Challenges to Doing Digital Humanities in the Library." *Journal of Library Administration*, 53.1: 43-52, https://doi.org/10.1080/01930826.2013.756694 (accessed 19 Nov 2021).

Sayers, Jentery. (2016). "Minimal Definitions." *Minimal Computing*, http://go-dh.github.io/mincomp/thoughts/2016/10/02/minimal-definitions (accessed 18 Nov 2021).

Thylstrup, Nanna Bonde. (2018). *The Politics of Mass Digitization*. Cambridge, MA: MIT Press.

Visual Media Workshop at the University of Pittsburgh. (2021). *The Socio-Technical Sustainability Roadmap*, http://sustainingdh.net (accessed 22 Nov 2021).

Temporal Topologies: Inflecting the telling and the told of historical narratives

Drucker, Johanna
drucker@gseis.ucla.edu
University of California, Los Angeles, USA

Dörk, Marian
marian.doerk@fh-potsdam.de
Fachhochschule Potsdam, Germany

Morini, Francesca
francesca.morini@sh.se
Fachhochschule Potsdam, Germany; Södertörn University, Södertörn, Sweden

LaCelle-Peterson, Nathaniel
nlapeterson@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Berlin, Germany

Rinderlin, Jonas
Jonas.Rinderlin@hkw.de
Haus der Kulturen der Welt, Berlin, Germany

Barker, Elton
elton.barker@open.ac.uk
The Open University, Milton Keynes, UK

Rosol, Christoph
rosol@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Berlin, Germany; Haus der Kulturen der Welt, Berlin, Germany

Wintergruen, Dirk
dwinter@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Berlin, Germany

The challenge of creating interactive graphical expressions for the representation of temporality in humanistic documents has spawned multiple digital humanities projects over the last decades (e.g., Temporal Modeling Project, PeriodO, Narrelations). These are premised on a distinction between positivistic temporal models that take time references as givens and those that understand temporal objects as discursive representations. The Temporal Topologies collaboration supports humanistic, hermeneutic approaches to temporality and chronology by creating graphical tools to create, display, analyze, and interpret temporal models while also using narratological frameworks of analysis. The project includes design of a pipeline that transforms semantically enriched HTML markup into JSON to generate graphic display, and a prototype for direct editing and argument production. Temporal models are the core case studies in this panel, but the larger research goal is to make it possible to attach arguments to specific data elements through a dynamic interface. We use the concept of "inflection" to indicate alterations of a display coupled with the underlying data. Rather than existing interaction techniques provided for data visualization (such as filtering elements or toggling layers), inflections are discursive interactions that are meant to be saved, shared, and shown. To realize this vision, interpretative capabilities need to be baked into the stack, from the visual interface to the data storage and back. The test bed for Temporal Topologies is the Anthropocene Curriculum of the Haus der Kulturen der Welt (HKW) and the Max Planck Institute for the History of Science (MPIWG) in conversation with connected initiatives at the University of California, Los Angeles (UCLA), Urban Complexity Lab, FH Potsdam (UCLAB/FHP), and The Open University (OU).

Johanna Drucker and Marian Dörk: Overview and the concept of "inflection"

This presentation describes the context for the Temporal Topologies project. We explain the rationale for an interactive interface linked to data that can support production of humanistic arguments about temporal phenomena and introduce the concept of "inflections" central to this endeavor. We propose the modification of the OWL Time ontology to accommodate discursive approaches to the representation of temporality. This paper frames the multiple facets of this work and the aim of provoking a conversation about its applicability to other projects in digital humanities. The panelists will focus on the conceptual development and technical design process for each component: markup, display, narratological approaches, date-specific Time Reference Systems, the relative temporal logic of James Allen, explicit and implicit temporality, varying time scales, and symbolic time schemes.

Christoph Rosol: Temporal Topologies in the Anthropocene Curriculum

"Every historical era is likewise multi-temporal, simultaneously drawing from the past, the contemporary, and the futuristic. … A circumstance is thus polychronic, multi-temporal, revealing a time that is gathered together, with multiple pleats." For the late philosopher and science historian Michel Serres (historical) time was a topological space, a folded and pantopic model in which temporally very distant instances close ranks, while others, chronologically more proximal ones, are separated. In the Anthropocene, the crisis-laden geological epoch upon us dominated by the effects of human impact on the Earth system, we see not only human and Earth history folding into one another but a tight entanglement and interaction between multiple temporal scales, causes and effects, sinks and sources. The Anthropocene necessitates a reorientation of intellectual comprehension, especially with regard to temporal horizons and their critical interconnectedness. The Anthropocene Curriculum (AC) initiative by HKW and MPIWG is an ambitious, long-term attempt to generate and foster a topology of Anthropocene knowledge. A topology describes the world qualitatively. As a science of connections, relationships, and mutual interdependencies, this Anthropocene topology is drawn from vernacular studies and ground-based encounters that the global AC network has elaborated over the last years. With the Temporal Topologies project, meant not least as a tribute and proof-of-concept of the Serresian model of topological time, the AC is now engaged in the development of a novel visual model of the qualitative temporal data.

Jonas Rinderlin and Nathaniel LaCelle-Peterson: Data and date tagging in the Anthropocene Curriculum project

The Anthropocene Curriculum's vast compilation of essays, reflections, field notes, geological data, and media materials makes it an ideal testing ground for new graphical

tools for the creation, analysis, and comparison of temporal models. Making the temporal information contained in such a diverse corpus accessible as structured data that can be processed using computational methods comes with its own set of challenges. While existing ontological frameworks, like the Time Ontology in OWL, provide a sophisticated means to describe temporal properties, the positivist structure of their model is insufficient for humanistic inquiry. In the context of the AC the specific question that arises is how to adequately represent the imbrication of multiple time scales. Drawing on distinctions from narratology and previous work by Evelyn Gius (heureCLÉA), we identified a small but crucial set of enhancements to the Time Ontology in OWL that allow us to incorporate discursive aspects into our own approach. In making additions to OWL Time, we took the humanistic context, the collaborative process, and practical requirements into account: new information about the narration of time is structured alongside existing OWL categories, and open fields allow for more flexible project-level labeling practices.

Francesca Morini: Visualizing temporal relations within conflicting discourse

This presentation describes the design process developed to render visible the relationalities among temporal events within discourse, and to visualize interpretative work performed by researchers. This visualization is designed to support practitioners in making sense of similar yet temporally conflicting narratives, as well as in discussing and generating new inflections. Despite the various contributions at the intersection of literary studies, narratology and interface design, graphical representations of narratives seldom support interpretative interactivity with text. Scholars rely on a limited set of tasks, which do not alter the way the underlying data are structured or represented. Moreover, narratives are rarely compressed into meaningful overviews capable of enabling comparison and rendering temporal conflicts visible. In the context of the Anthropocene Curriculum project, we devised temporal signatures as graphical representations of narratives that render discursive and temporal relationality visible. These visualizations are designed to foster critical understanding of temporally conflicting narratives, through techniques that enable transitions, synoptic views, and comparison across scales and models. Additionally, through the cooperation with the Heterochronologies research at UCLA, we provide new ways of extending editing powers to contest narratives and temporal structures, while keeping a trace of such disputations.

Dirk Wintergruen: Modelling socio-epistemic networks

ModelSEN is a research project funded by the German Federal Ministry of Education and Research. It aims to find ways to describe and analyze the dynamics of knowledge systems by applying different graph and network theory models, including social network analysis and agent-based modelling. An essential part of this project is the integration of different knowledge bases through semantic modelling of data produced by different data providers. Of course, modelling temporality is a central part of creating an integrative model, not only for modelling the data input. Rather, it is about semantically encoding the results of different models that simulate the dynamics of knowledge systems to make them comparable. Furthermore, visualization techniques are necessary to communicate the results and make them comparable for a wider audience that is not used to interpreting numerical data. But also, to identify patterns that often only become visible when the data can be visualized and interactively manipulated.

Elton Barker: Exploring time and space in a non-modern narrative

This presentation tackles a particularly challenging example of temporal mapping — Pausanias's first-century CE Periegesis Hellados (Description of Greece). A ten-book journey around Greece, the Periegesis is as much modelled on time as it is on space, as Pausanias records stories about the locations, buildings and objects through which he moves. Mapping the dynamic relationship between space and time (chronotope: Bakhtin 1981) presents a significant challenge to digital visualisation, especially in texts that are also about time, where temporal representations sometimes align with narrative time (chronological sequence) and sometimes do not (Ricoeur 1984). The challenge is all the greater in a non-modern author like Pausanias, for whom canonical dating (e.g. "490 BCE", "2021 CE": Feeney 2007) or periodizations (e.g., "archaic", "classical", etc.) risk misrepresenting the chronotopic dimension of his journey. Here I discuss theDigital Periegesis project, which is using the semantic annotation tool Recogito to encode place and time in Pausanias. As well as outlining our annotation practice, I discuss the different mechanisms by which time is described by proxies (people and events) and efforts to map them to global authorities (such as PeriodO or Wikidata). I also consider the potential affordances of annotating place and time for humanities Linked Open Data, as well as reflect on the design challenges for conducting temporal-spatial

analysis within this ecosystem (using the visualization tool Peripleo). I conclude with some reflections on how the work on the Anthropocene Curriculum can contribute to visualizing narrative temporality in a narrative as complex as Pausanias's.

Global Perspectives on Critical Infrastructure and the Digital Humanities Lab

Hannah, Matthew Nathan
hannah8@purdue.edu
Purdue University, United States of America

Connell, Sarah
sa.connell@northeastern.edu
Northeastern University, United States of America

Dodd, Maya
maya@flame.edu.in
FLAME University, India

Ope-Davies (Opeibi), Tunde
bopeibi@unilag.edu.ng
University of Lagos, Nigeria

Povroznik, Nadezhda
povroznik.ng@gmail.com
Perm State University, Russia

Rittenhouse, Brad
bcrittenhouse@gatech.edu
Georgia Technical University, United States of America

As Digital Humanities has evolved globally, the infrastructural demands for lab spaces have become of paramount importance. This panel applies a critical lens to the question of infrastructure for building and sustaining the DH lab. Each panelist represents a lab space with particular challenges and opportunities, and we hope to delineate key differences and similarities in the design, management, and mission of the DH lab as a theoretical and practical matter. Theorizing our labor as lab designers, managers, and directors requires an orientation toward infrastructure—a humanities infrastructure—which is an essential component of the digital humanities lab. This panel attends to various components of material and cyberinfrastructure which impact our lab spaces directly. Each panelist analyzes their space within very different contexts, but we hope to generate synergies across national, conceptual, and disciplinary boundaries, exposing both the tensions and similarities among our various spaces.

We see our particular cluster of infrastructural concerns as reflective of what Alan Liu, Urszula Pawlicka-Deger, and James Smithies (2021) articulate as the burgeoning field of "critical infrastructure studies," which has "emerged as a framework for linking thought on the complex relations between society and its material structure." For those of us developing DH labs, such a theoretical orientation provides important optics for assessing the DH lab model we have inherited. In this sense, "critical" refers both to the indispensable systems we rely on to maintain labs and, at the same time, the evaluative and reflexive perspective we hope to develop through dialogue with one another and with conference attendees. In our experience, the DH infrastructures we support reflect Sheila Anderson's (2013) call to "view infrastructure as a material and experiential presence that is embedded in the practices and experience of research," and each panelist theorizes this presence within and across the spaces represented on this panel.

Such attention to the material infrastructure animates our own work on the presence and persistence of the Digital Humanities lab as a physical space and social hub. In this account, the DH lab represents both unique infrastructural formations and deformations within the broader network of university resources and staff and, at the same time, an aggregate of inter- and multidisciplinary research, teaching, outreach, and collaboration. The particular nexus between the social and the material, which energizes the field of critical infrastructure studies, also vivifies our own thinking about the present—and future—of the DH lab. And as we confront the infrastructural instabilities of the post-COVID academy, we must critically assess and respond to worsening austerity.

Sarah Connell (Northeastern University)

Critical Infrastructures for Conscientious Work: A Case Study

Northeastern's NULab for Texts, Maps, and Networks currently comprises 41 faculty members, seven graduate Fellows, two faculty co-directors, and one half-time staff position. Most of the large-scale coordination of the NULab falls to the staff role, but, as these numbers suggest, much of the local organization and day-to-day work of the NULab is done by the graduate Fellows. This paper will share strategies for building effective infrastructures in labs that rely primarily on graduate labor—and will also discuss

some challenges that remain. Lab infrastructures and their labor models are a pressing concern in DH, as indicated by the fact that the "laboratory turn" and invisible labor have been the focus of two recent special issues in *Digital Humanities Quarterly* (Graban et. al., 2019; Oiva and Pawlicka-Deger, 2020). Critical attention to infrastructures can mitigate the exploitative aspects of graduate labor while ensuring that the lab itself runs smoothly. The NULab focuses considerable effort on training and documentation, following a model of social knowledge creation in which Fellows are equal partners in planning and decision-making. There are challenges associated with this model that even the best-designed infrastructures cannot eliminate: for instance, Fellows are always positioned as students first and employees second, which means that lab operations sometimes need to be recalibrated. This paper will share the specific tactics the NULab has developed, consider how these might be adapted in other contexts, and take up some of the questions that large-scale reliance on graduate labor raise for the digital humanities as a field.

Maya Dodd (FLAME University)

"What's Missing?" : On Critical Digital Infrastructures in India

While it has been noted that infrastructures are central to the practice of digital humanities, it is also true that limits to the digital make infrastructures for DH labs in India particularly fragile. As anthropologist Akhil Gupta (2018) states, infrastructures "are a process [not a thing] that is characterized by multiple temporalities [and] open futures." Affording DH/dh project work in India is often a function of both imagination and infrastructure. The structural exclusion of the non-English speaking is a defining impediment to DH labs in India, and we see how this frames institutional possibilities of curation and distribution. Mostly, extant usage of digital tools rests on the overall systemic conception of access, via English. The need to develop infrastructures across several Indian languages and to examine the need for open access resources (such as virtual labs) might offer some possibilities to combat existent shortages. For DH labs to scale up in India, an exploration of what is yet possible would also need to contend with historical barriers that stand out here. To name some, 1. The barrier to accessing higher education in languages beyond English that structures the research ecosystem, 2. The fact that access to an indigenous publishing system with reach and inter-operable legitimacy and use is absent, and 3. The historical impediments for both students and faculty to global access, funding and exposure (due to expense). Since neither connectivity nor robust funding can be assumed, even in the formal education settings of Indian Universities, to imagine DH labs, tools and resources in India also necessitates the consideration of offiine techniques. In India, digital affordances need to be imagined beyond known DH lab infrastructures of the global North.

Matthew N. Hannah (Purdue University)

Building Towards a Humanities Infrastructure

Building a sustainable infrastructure for Digital Humanities at a predominantly STEM university presents unique critical challenges in retaining an emphasis on transdisciplinary and interdisciplinary research while, at the same time, securing space for a humanities that is potentially undervalued. In building a space, one must consider the critical mission of the lab in advancing and supporting both digital scholarship in general and humanities scholarship in particular. As James Smithies (2017) points out, the humanities have already developed extensive and global infrastructure: "There is strong reason to argue that the humanities already have a larger cyberinfrastructure than the science and technology communities, one that is more global, more connected, and more complex in both technical and epistemological terms." But how might this humanities infrastructure develop vis-a-vis existing STEM infrastructure in the development of Digital Humanities? Are there potential overlaps between the two and what are the potential drawbacks to redeploying STEM resources for humanities projects? This paper maps the disciplinary topography of existing infrastructure at a predominately STEM institution and provides strategies for leveraging such resources for explicitly DH spaces and projects. Despite the paucity of material resources for DH within certain zones of the university, the presence of neutral spaces within the Libraries and School of Information Studies have enabled fruitful transdisciplinary collaborations, which have extended to the development of unique mixed-use spaces.

Tunde Ope-Davies (Opeibi) (University of Lagos)

The Role of Infrastructure and the Future of DH Labs in Emerging Spaces

As the growth of Digital Humanities (DH) continues to generate greater excitement among scholars across the globe, the provision and sustenance of critical infrastructure to drive and escalate this momentum appears to constitute some concerning component. It has thus become necessary to establish a more pragmatic approach between theoretical optimism and practical reality in order to reassess the

challenges confronting DH Labs especially in emerging communities of practice. This presentation therefore examines the existing and inescapable systemic tension among the various interacting variables in this space such as the availability of material infrastructure, institutional support, human capital resources, capacity building initiatives, affordable relevant technologies and social forces which potentially impact the growth of DH initiatives in these emerging spaces, with more focus on Nigeria and the sub-region. The paper thus speaks to the key role of critical infrastructure in promoting and sustaining DH Labs that may [in]-directly influence the potential transformative power of DH beyond the fields of humanities.

Brad Rittenhouse (Georgia Technical University)

Diversity and Sustainability through Infrastructure

Laboratories are often culturally gendered and racialized spaces associated with competitive and exhibitionist performances of technical "expertise," assigned uninterrogated value as sites of production and innovation, and operate with clique membership. These constructs serve as barriers to a range of potential lab participants who may not recognize themselves or their interests in the generic image of a lab. The tech world has traditionally been set up by white men in ways that ensure (mostly) white men succeed. Moya Bailey (2011) details this problem in DH, noting that the field tends toward whiteness and maleness, requiring structural change beyond the "add and stir" method of adding diversity. In the presentation, Rittenhouse will speak about his experiences using lab infrastructure and other strategies to increase DEI in a research and service lab setting. The presentation is based on experience as a lab manager of a grant-funded DH makerspace at an R1 technical institute, a setting which at times can exhibit many of the aforementioned DEI shortcomings. Rittenhouse will present institutional diversity numbers, which fall short of national and local demographics in many areas, and detail strategies to nearly double diversity figures in most major categories over the past five years. These efforts include intentionality in funding and supporting projects from diverse researchers, incorporating inclusive language and practices into hiring and recruitment efforts, and planning events and outreach focusing on issues of diversity to sincerely and meaningfully transform the community of a lab to one that is more inclusive, equitable and, ultimately, sustainable.

Nadezhda Povroznik (Perm State University)

Growing DH Center: From Zero to Shared Infrastructure

The Center for Digital Humanities at Perm University was established in 2016 based on the Laboratory of Historical and Political Informatics and its rich background. Various projects implemented in the Laboratory, including history-oriented systems (Kornienko, Gagarina, and Povroznik, 2021), could benefit from digital infrastructures for optimization of research tasks and processes. The Center's activities have not been related to investing in the development of digital infrastructure for a long time, since most of the projects have been grounded on open services or free platforms for educational organizations or cultural heritage institutions. For example, SketchFab is such a platform that has been used by the Center for publishing 3D models of digitized objects from the collection of the Museum of History. The specificity of the Center has become the growing interdisciplinarity of projects. In 2020, the Center became part of the ARTEST project. One of the tasks of the project is to create a virtual laboratory, a shared space for co-creation activities during the courses of Master Programs in Digital Humanities. Another contemporary trend is convergence and networking with the other labs and centers within the university. The Perm University develops *Priority 2030*, a Strategic Academic Leadership Program and the shared digital infrastructure for the humanities and "hard" sciences is under discussion. During the panel, the issues of the acceptance, adaptation, and choosing the appropriate models of the digital infrastructure for the DH center's growth will be discussed.

Bibliography

Anderson, S. (2013). What are research infrastructures? *International Journal of Humanities & Arts Computing: A Journal of Digital Humanities* 7 (1/2): 4–23. doi:10.3366/ijhac.2013.0078.

ARTEST project. Virtual Platform. URL: http://artestproject.com/virtual-platform (accessed 20 April 2022).

Bailey, M. (2011). All the digital humanists are white, all the nerds are men, but some of us are brave." *Journal of Digital Humanities* 1(1). http://journalofdigitalhumanities.org/1-1/all-the-digital-humanists-are-white-all-the-nerds-are-men-but-some-of-us-are-bra ve-by-moya-z-bailey/ (accessed December 6, 2021).

Graban, T., Marty, P., Romano, A., and Vandegrift, M. (eds.) (2019). Invisible work in digital humanities. *Digital Humanities Quarterly*. 13.2. 2019.

Gupta, A. (2018). The future in ruins. *The Promise of Infrastructure*. Anand, N., Gupta, A., and Appel, H. (eds.). Durham: Duke University Press.

Kornienko S., Gagarina D., Povroznik N. (2021). Historical information systems: Theory and practice. HSE Publishing House.

Liu, A., Pawlicka-Deger, U., and Smithies, J. (2021). CFP: Critical Infrastructure Studies, https://dhdebates.gc.cuny.edu/page/cfp-critical-infrastructure-studies-digital-humanities (accessed 20 April 2022).

Oiva, M. and Pawlicka-Deger, U. (2020). Lab and slack: Situated research practices in digital humanities." *Digital Humanities Quarterly*. 14.3.

Smithies, J. (2017). *The Digital Humanities and the Digital Modern*, Palgrave Macmillan UK,

The Ethical Considerations of Diverse DH Pedagogy

Licastro, Amanda Marie
amanda.licastro@gmail.com
University of Pennsylvania, United States of America

Stringfield, Ravynn K.
rkstringfield@email.wm.edu
William and Mary University, United States of America

Earhart, Amy
aearhart@tamu.edu
Texas A&M University, United States of America

Losh, Elizabeth
lizlosh@wm.edu
William and Mary University, United States of America

This session on digital humanities pedagogy will offer practical pointers for scaffolding courses around DH projects with attention to the most difficult issues including:

In addition to discussing the pedagogical principles in four model courses designed by the presenters – which range from undergraduate general education courses to special topics graduate seminars – this session will also provide an annotated bibliography of useful DH readings about pedagogy including works that are part of larger monographs (Risam 2019, D'Ignazio and Klein 2020, etc.), edited collections such as Digital Pedagogy in the Humanities and Critical Digital Pedagogy, and journals such as Hybrid Pedagogy and The Journal of Interactive Technology and Pedagogy.

Exploring Immigrant Narratives through Extended Reality

Speaker one will present on a first-year literature course, certified as fulfilling "Intercultural Knowledge Competency" (IKC), focused on literature written by and about those seeking citizenship in the United States. The students in the course consider poetry, short stories, novels, and cultural objects such as music videos, viral content, and news publications as well as Virtual Reality applications that raise questions about the language and imagery used to define immigrants, refugees, and asylum seekers. Using narrative as a means of cultivating empathy, and service learning partnerships to encourage long-term action, this course employs what Lisa Blankenship coined "rhetorical empathy" to structure research and reflection. The introduction and critique of the VR applications is intended to test the hypothesis that immersive experiences can teach empathy, a premise currently under debate in both the humanities and social sciences. In considering the criticism of scholars such as Lisa Nakamura and Liz Losh alongside the advocacy of artists such as Nony de la Pena and Gabo Aurora, the speaker will demonstrate ways of pairing XR content with traditional creative and academic texts to provoke respectful conversation, sustained service and activism within the community, and academic research projects.

Fantasy New Media Narratives

Speaker two will discuss an upper-division undergraduate course that expands the field to recognize more potential creators of digital projects and takes a future-oriented vision for Black digital humanities. It explores the concept of #BlackGirlMagic and what it has meant for Black women to create new worlds and alternate realities for Black woman- and girlhoods to exist. She will discuss how weaving together representations of Black women and girls feature in fantastic, digital and futuristic media—including literature, comics, music, film, and webseries—gave students a creative opportunity to: (1) to familiarize students with the fields of Afrofuturism, the Black Fantastic, and Black Speculative Fiction, etc. (2) to introduce students to prominent Black women artists and creators and their works, and (3) to explore how we conceptualize and construct Black woman and girlhood through the lens of Fantastic, Futuristic and Digital creative and scholarly work.

In crafting a course centering Afrofuturism, The Black (and/or Dark) Fantastic, and Black Speculative Fiction through a variety of media, the speaker aimed to provide a classroom space in which students were able to articulate how Black woman and/or girlhood is constructed with the use of magic as a tool or a lens in written assignments and in a final (creative) project of their choice, developing their identities as scholars, creators and fans. The speaker will argue that offering rigorous feedback and support of individually and collaboratively constructed creative final projects was a necessary component of building a classroom safe for intellectual inquiry and to practice tenets of speculative art, which centers imaginative co-creation as an integral first step toward new futures.

Ethical strategies for integrating "hard history" digital projects into the classroom

Speaker three will present on a graduate course designed to explore Black Digital Humanities, the intersections of critical race studies, African-American literature, editing and recovery, and digital humanities. Important to our local environment, the paper will discuss how to create a specialty class that invites students to participate that are not specialists in digital humanities nor Black studies. Following the concept of "juxtaposition," what Kelly Baker Josephs and Roopika Risam see as the intersection "of disciplines, cultures, and methods" (Introduction, The Digital Black Atlantic 2021), the course consists of traditional class discussion with regular lab days that bridge the theoretical with methodological and applied contents. Students learn about Black dh through engagement with a local Black dh project, the Millican Massacre, 1868, a digital project that is recovering a race massacre. Discussing ways that theory and practice are integrated within the course, particularly for students that do not necessarily have any digital humanities skills, the paper will provide samples of lab activities and group projects. The balance of theory and practice is embedded in class design, and the paper will address the tradeoffs in such work. Further, the project has community connections, and students learn about ways that public humanities projects ethically work with sensitive histories that continue to impact communities.

Facilitating Agency and Criticality to Support More Diverse Cohorts [8]

Speaker four will discuss pedagogy for graduate students with a focus on how digital humanities methods and theories can counter self-selection biases toward those who already identify as technically literate. This speaker will discuss how PhD candidates from American Studies, History, and Anthropology can be given more agency to develop their own DH methods and make choices about curating their own data sets. Rather than reproduce a specific lab's practices, students can be exposed to a broader range of ways to annotate data sets, select among multiple platforms for online exhibitions or the composition of scholarly hypertexts, and choose among various tools for mapping social networks, incorporating GIS, and structuring linguistic corpora and media archives. At the same time, hands-on sessions should also be embedded in a larger context of rich discussions with DH practitioners about the ethical dilemmas and institutional challenges that come with taking responsibility for shaping what Roopika Risam has called the "digital cultural record." Theoretical readings may be drawn from feminist, Black, postcolonial, and queer DH scholarship that show how digital humanities projects can also promote surveillance, reinscribe ideologies of white supremacy, and normalize social sorting and exclusionary logics. In other words, this pedagogical approach goes beyond simply modeling best practices for reaching new publics or formulating new research questions to include students as stakeholders in critical processes of the review of scholarship that show why some digital humanities projects might raise troubling questions about control, consent, credit, and cooperation.

Digital debating cultures: Communicative practices on Reddit

Messerli, Thomas C.
thomas.messerli@unibas.ch
University of Basel, Switzerland

Dayter, Daria
daria.dayter@tuni.fi
Tampere University, Finland

Bohmann, Axel
axel.bohmann@anglistik.uni-freiburg.de
University of Freiburg, Germany

Donlan, Lisa
lisa.donlan@manchester.ac.uk
University of Manchester, UK

Maccori Kozma, Gustavo

gustavo.maccori.kozma@saturn.uni-freiburg.de
University of Freiburg, Germany

Leuckert, Sven
sven.leuckert@tu-dresden.de
TU Dresden, Germany

Liimatta, Aatu
aatu.liimatta@helsinki.fi
University of Helsinki, Finland

Mahler, Hanna
mahlerhanna@aol.de
University of Freiburg, Germany

Massanari, Adrienne
adrienne@american.edu
American University, Washington D.C., USA

McConnell, Kyla
kyla.mcconnel@anglistik.uni-freiburg.de
University of Freiburg, Germany

Tosin, Rafaela
rafaela.tosin@anglistik.uni-freiburg.de
University of Freiburg, Germany

Panel abstract

The social media platform Reddit understands itself as a “home to thousands of communities”, where every used can find their community (https://www.redditinc.com). As researchers in humanities, we find that the submissions and comments posted to Reddit’s subreddits do indeed comprise authentic digital human interaction by groups of people that are in some cases prototypical communities and in other cases merely chance encounters of users who find themselves oriented towards the same virtual space. The collective communicative acts of Reddit users can be positioned in the tradition of computer-mediated communication (CMC) – as one key site of digitalised communication, shaped partially by the affordances provided by the platform, and uniquely available to researchers not just in terms of their linguistic content, but also their multimodal context and discursive structure. Importantly, however, Reddit (sub-)communities are not necessarily subject to identical communicative patterns – within each community, user types and even individual users communicate following particular patterns or even idiosyncratically.

Our recently formed interdisciplinary network, copRe (**co**mmunicative **p**ractices on **Re**ddit – copre.org), is dedicated to exploring Reddit discourse(s) from different theoretical perspectives, but all with the aim to contribute to the understanding of Reddit’s own communicative culture as well as the exploration of digital practices more generally.

Specifically, our panel at the DH2022 conference explores aspects of digital culture and participatory culture, manifest in the communicative acts of different (sub-)communities on the online social platform Reddit. These subreddit-communities and the digital genres they give rise to are sites of linguistic innovation as well as of new debating practices – from the combative far-right subreddit r/The_Donald to the more harmonious r/changemyview. They let us gain insights into individual and group identities, as on r/Mountaineering, and they raise question of methodology, such as the understanding of text length as both a challenge for research and a motivated choice of text authors and the employment of mixed-methods to gain insights that are both driven by big data as well as by in-depth understanding of individual and collective communicative acts.

Language innovation and diffusion online.

Lisa Donlan (University of Manchester)

Who are the innovators of lexical terms online? What are the community roles of the early adopters who successfully diffuse linguistic innovations?

In offline communities, the weak-tie theory of language change envisions the innovators of linguistic forms as peripheral to a community while early adopters are the community's central members. However, the only study to explore the applicability of the theory in an online Community of Practice (CoFP) was grounded in an unusual linguistic context. My research addresses this gap in the literature by using a mixed-methods approach to analyse the status of the innovators and early adopters of four community-salient innovative linguistic forms which diffused through an online music-orientated CofP, Popheads.

Contrary to expectations, three of the four forms studied were innovated by non-peripheral members who scored highly across multiple markers of status. This departure from previous findings may be related to the fact that linguistic creativity is highly valued in many virtual contexts. Consequently, high-status members may perceive linguistic innovation as desirable behaviour online.

This research also found that identifying the hierarchical structures that underpin a community leads to more precise descriptions of the characteristics of early adopters. Specifically, it has been possible to conclude that early

adopters are prolific contributors, whose posts are successful at generating discussion, and who are on inbound trajectories in the community. Therefore, to speak of an early-adopter as being 'central' or 'high-status' is, I argue, ultimately too vague and fails to acknowledge the multidimensional nature of status.

Functions of text length on Reddit

Aatu Liimatta (University of Helsinki)

In corpus-linguistic studies, text length is typically seen as a potential confounding factor (see e.g. Liimatta, 2020), largely because its effects have been difficult to study using even the largest traditional corpora. However, like any other linguistic choice, the length of a text is also a choice made by the writer or speaker: it is also affected by the communicative purpose of the text and the limitations and affordances of the communicative situation.

Fortunately, large social media datasets with a range of text lengths have allowed us to approach this previously unassailable topic. Reddit is particularly interesting in terms of text length, since the length of a Reddit comment is free to vary according to the commenter's needs. Recent studies have shown that Reddit comment length is linked to the distribution of functional linguistic features: for instance, simple information-seeking comments tend to be very short, whereas narrative registers appear to favor longer comments on average (Liimatta, forthc.).

In order to further explore the role and functions of comment length on Reddit, I analyze a number of subreddits in terms of both the distribution of comment lengths and the distribution of functional linguistic features across comment lengths. To do this, I make use of a large-scale dataset of Reddit comments and a simple but powerful pooling-based computational methodology.

Register variation in Reddit comments - A multidimensional analysis

Hanna Mahler, Kyla McConnell, Axel Bohmann, Gustavo Maccori Kozma, Rafaela Tosin (University of Freiburg)

Researchers are increasingly becoming interested in the many opportunities that Reddit provides for linguistic analysis. In this large-scale natural language processing project, we focus on register variation within Reddit comments (inspired by Liimatta 2016, 2020).

We analyze Biber's (1998) linguistic features for register analysis, as well as platform-specific features, on all Reddit comments since 2005, using the Pushshift Reddit Corpus (Baumgartner et al. 2020). We are using this feature annotation to implement a short-text MDA (Clarke & Grieve 2019), a version of Biber's (1988) multi-dimensional analysis, to find out which dimensions describe the linguistic variation found on the platform and whether the topical "subreddits" can be described as different registers. Our method also promises to serve as a useful tool for analysing other topics such as adaptation of linguistic norms or register diversification over time.

Our study therefore adds to the state of knowledge in several ways:

1. We regard a single comment as one text (with features extracted on the sentence level), which allows us to accurately locate linguistic variation within individual users.
2. We train a tagger specifically to overcome previous difficulties of tagging social media data (e.g. Banga & Mehndiratta 2017), based on data from Behzad & Zeldes (2020) and Gessler et al. (2020).
3. The feature extraction script, a refined and elaborated version of Biber's (1988) initial features, is written in Python and will be made openly available.
4. Our long-term goal is to develop an MDA solution that captures variation within and among all (English) subreddits.

Combating the Far-/Alt-Right on Reddit: Lessons from r/AgainstHateSubreddits

Adrienne Massanari (American University, Washington D.C.)

Reddit embodies a carnivalesque spirit, often reflecting a kind of geek masculinity (Kendall, 2011) that champions both niche, technical prowess and clever humor (Massanari, 2015). At the same time, communities engaging in far-right rhetoric, such as the now-banned (and widely popular) r/The_Donald, have flourished in part because the platform relies almost exclusively on volunteer labor to moderate and grow communities (Matias, 2019). Shifting the responsibility and risk of moderating onto unpaid individuals allows Reddit to remain a "lean" organization with few employees, but also creates a kind of plausible deniability when it comes to so-called "alt-right" subreddits.

In response to the growing threat that these subreddits present, and the lack of response from Reddit administrators, activists on the platform have created their own communities focused on highlighting hate speech pervasive on the platform. One such example is r/AgainstHateSubreddits, which is dedicated to exposing subreddits that may superficially conform to Reddit's few rules, but also engage in transphobic, misogynistic, Islamophobic, and racist rhetoric. Through a critical discourse analysis (Fairclough, 2013) of popular postings on the subreddit, I explore how this community challenges

Reddit's politics and offers an ethical counterpoint to the toxic geek masculinity that pervades much of the platform. Drawing on work from platform studies (Bucher, 2018; Gillespie, 2010) and design justice (Costanza-Chock, 2020; D'Ignazio & Klein, 2020), I argue that Reddit's governance, design, and platform policies work implicitly welcome and mainstream far-right communities, but that spaces like r/AgainstHateSubreddits provide critical forms of resistance and community for activists.

Share my view: Harmonious debating culture on r/changemyview

Thomas C. Messerli (University of Basel), Daria Dayter (University of Tampere)

In current times, digital discourses are often understood in terms of polarization. Public lay metadiscourses are full of references to social bubbles and disparate parts of society, whereas academic scholars give a lot of focus to binary categories such as information/disinformation, truth/post-truth or outrage culture. Within this context, the debating culture on the subreddit r/ChangeMyView (CMV) stands out because it encourages what we could term persuasibility – the capacity or willingness of someone to change their opinion when encountering new information. While some work has been done on the specific strategies that commenters use to achieve the task at hand, i.e. to change the original poster's (OP) view, little attention has been paid to the question how prepared OPs actually are to change their mind and how this "malleability of opinion" (Tan et al. 2016: 621) is discursively constructed. From this perspective, original posts – *submissions* in Reddit terminology – are firstly performances of persuasibility, and secondly access points to persuasible-persuasive pairings, in which the subreddit community enacts its codified and tacit norms. In order to explore these pairings, we make use of the CMV corpus we have compiled and specifically compare submissions, delta-awarded comments, i.e. those comments that have changed the OP's view, and the OP's responses to delta-awarded comments. We do this comparison itself with a mixed-methods approach that is grounded in qualitative annotation of persuasibility in a sample of r/changemyview threads and scaled up to the corpus using corpus linguistic methods.

"Science has no business in the mountains": Stance-taking and expert knowledge on r/Mountaineering

Sven Leuckert (TU Dresden)

Stance-taking, as popularised in pragmatics and sociolinguistics by Du Bois (2007), refers to "the speaker's (or writer)'s relationship to (a) the topic of discussion, (b) the interlocutor or audience, and (c) the talk (or writing) itself" (Kiesling et al. 2018: 684). On social media, stance-taking plays an important role in the discursive construction of relationships and may be employed as a gatekeeping device. In this talk, I focus on strategies of stance-taking as it is linked to the expression of expert knowledge on the subreddit r/Mountaineering. On this subreddit, stance-taking represents a dominant tool to establish who can be considered an expert and, hence, part of the knowledgeable in-group.

In this talk, I explore which specific linguistic phenomena are employed by users of the subreddit to express stance in situations where expertise in mountaineering is in focus. After an initial manual assessment of recurring phenomena on the basis of randomly selected threads, a quantitative approach inspired by Kiesling et al.'s (2018) annotation scheme is used to establish the bigger picture of how stance-taking is employed as a gatekeeping device on r/Mountaineering. For this study, the entirety of r/Mountaineering from 2012 to August 2021 has been scraped and is taken into consideration. In sum, the findings suggest that, while quantitative methods are a useful addition in the investigation of stance on Reddit, they can only be complementary to an in-depth study of stance-taking phenomena in their discursive context.

Bibliography

Banga, R. and Mehndiratta, P. (2017). Tagging Efficiency Analysis on Part of Speech Taggers: International Conference on Information Technology (ICIT: 264–267.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M. and Blackburn, J. (2020). The Pushshift Reddit Dataset: Proceedings of the International AAAI Conference on Web and Social Media, 14th edn.

Behzad, S. and Zeldes, A. (2020). A Cross-Genre Ensemble Approach to Robust Reddit Part of Speech Tagging. http://arxiv.org/pdf/2004.14312v1.

Biber, D. (1988). Variation Across Speech and Writing. Cambridge University Press.

Bucher, T. (2018). If...then : algorithmic power and politics. Oxford University Press.

Clarke, I. and Grieve, J. (2019). Stylistic Variation on the Donald Trump Twitter Account: A Linguistic Analysis of Tweets Posted between 2009 and 2018. PloS one, 14(9), e0222062.

Costanza-Chock, S. (2020). Design justice: Community-led practices to build the worlds we need. The MIT Press.

D'Ignazio, C. and Klein, L. F. (2020). Data feminism . The MIT Press.

Dayter, D, & Messerli, T. C. (2021). Persuasive language and features of formality on the r/ChangeMyView subreddit. Internet Pragmatics, 5(1): 165–195. https://doi.org/10.1075/ip.00072.day

Du Bois, J. W. (2007). The stance triangle. In Engelbretson, R. (Ed.), Stancetaking in Discourse. John Benjamins, pp. 139–182

Fairclough, N. (2013). Critical discourse analysis: The critical study of language. Routledge.

Gessler, L., Peng, S., Liu, Y., Zhu, Y., Behzad, S. and Zeldes, A. (2020). AMALGUM – A Free, Balanced, Multilayer English Web Corpus: Proceedings of The 12th Language Resources and Evaluation Conference: 5267–5275.

Gillespie, T. (2010). The politics of 'platforms'. New Media & Society, 12(3): 347–364.

Kendall, L. (2011). 'White and nerdy': Computers, race, and the nerd stereotype. The Journal of Popular Culture, 44(3): 505–524.

Kiesling, S. F, Pavalanathan, U., Fitzpatrick, J., Han, X. and Eisenstein, J. (2018). Interactional Stancetaking in Online Forums. Computational Linguistics 44(4): 683-718

Liimatta, A. (2016). Exploring Register Variation on Reddit: A Mulit-Dimensional Study . Master thesis, University of Helsinki.

Liimatta, A. (2020). Using lengthwise scaling to compare feature frequencies across text lengths on Reddit. In Rüdiger, S. and Dayter, D. (Eds.), Corpus Approaches to Social Media . John Benjamins, pp. 111–130.

Liimatta, A. (forthc.). Register variation across text lengths: Evidence from social media. International Journal of Corpus Linguistics.

Massanari, A. L. (2015). Participatory culture, community, and play: Learning from reddit . Peter Lang.

Matias, J. N. (2019). The Civic Labor of Volunteer Moderators Online. Social Media + Society, 5(2). https://doi.org/10.1177/2056305119836778

Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C. and Lee, L. (2016). Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions: Proceedings of the 25th International Conference on World Wide Web: 613–624. https://doi.org/10.1145/2872427.2883081

SpokenWeb: Curating Literary Sound in a Digital Environment

O'Driscoll, Michael
mo@ualberta.ca
University of Alberta, Canada

Luyk, Sean
sean.luyk@ualberta.ca
University of Alberta, Canada

Kroon, Ariel
kroon@ualberta.ca
University of Alberta, Canada

Morrison, Zachary
zmorriso@ualberta.ca
University of Alberta, Canada

Ambarani, Tejas
tejas.ambarani@ualberta.ca
University of Alberta, Canada

Miya, Chelsea
cmiya@ualberta.ca
University of Alberta, Canada

1. The SpokenWeb Project: A Brief Overview:

Michael O'Driscoll, Professor, Department of English and Film Studies, UAlberta

This panel proposes to draw from the experience of the SpokenWeb Project team at the University of Alberta, an institutional partner in an exciting multidisciplinary research consortium now in the fourth year of seven-year SSHRC Partnership Grant. The project seeks to establish a networked archive of digital, literary audio drawn from the institutional and community collections managed by partner universities and academic libraries across Canada. The collections comprise thousands of analogue recordings (primarily reel to reel and cassette tapes) dating from the 1960s to the 1990s that, in an aggregated, searchable digital form, constitute a valuable cultural, research, and learning resource. In the vast majority of cases, these audio objects have rested unheard and uncatalogued in dusty corners and are only now meeting their future audience imagined into existence some half a century ago.

Integral to the project is the development of protocols not only for preservation and access that will encourage research and pedagogical innovations in this developing field at the intersection of the digital humanities, library science, literary analysis, and sound studies, but also the establishment of fundamental principles and practices

for collections processing, metadata schemes, system interoperability, data analytics, and other technical matters necessary to the success of this innovative collaboration. SpokenWeb is also developing best practices in a multidisciplinary environment, ensuring high quality training and experiential learning for new scholars, and making inroads in community partnership development.

The University of Alberta SpokenWeb team is at a crucial juncture in the project's lifecycle: we are about to move from a focus on collection development to a focus on public access and community engagement. This diverse panel will speak from a range of disciplinary perspectives, and draws on both emerging and established scholars on the UAlberta SpokenWeb team. Central to our presentations is the question of how the curation of durational digital objects that feature human speech across a variety of audiotextual literary genres presents pressing questions and unique challenges in a digital environment. In order to address a representative range of issues, panelists will speak to matters of digital media collection development, background archival research, timestamping and the production of structural metadata, rights management and curatorial ethics, and designing a public collections portal. After several years of project development, we've learned much, including the fact that is a lot more to learn.

2. Sound Studies and the Repository Environment

Sean Luyk, Digital Projects Librarian, University of Alberta Library, Co-Investigator, SpokenWeb Alberta

Institutional repositories provide foundational infrastructure for digital humanities projects, including long-term preservation, the provision of (open) access, resource description, and intellectual property rights management. Projects that work to curate audiovisual content, however, present significant challenges for these scholarly communications services, as they are typically not designed with the affordances of durational media in mind. Using the SpokenWeb Alberta project as a case study, this presentation will discuss the use of institutional audiovisual repositories for sound studies projects, and the considerations and challenges involved for libraries and archives supporting this work. Topics discussed will include intellectual property rights management, collections workflows, audiovisual resource description, and media preservation, as they relate to the delivery of scholarly communications services for digital humanities projects.

3. Anomalies and Archaeology in the Archive: researching radio episodes

Ariel Kroon, PhD, SpokenWeb Research Assistant, University of Alberta

This section will introduce the audience to the archival research necessary to locate, confirm, and assure quality control for the audio being digitized. I will walk through the archiving process with focus first on a larger collection (UAlberta Archival radio show recordings, aired on campus radio network CKUA), then narrowing in on a single episode recording and detail the type of archaeological work required in order to identify its content and speakers, as well as the challenges of recording metadata for a new audio type that the project had not yet encountered, including the creation of appropriate language for metadata use, the labelling of weird and interesting speech acts, musical interludes, and surprising audio events. I will also discuss the challenges of tracking down related materials from a thirty-year-old radio show held in the collection of a library that no longer exists, which served a program (Radio and Television) that was shut down years ago.

4. Introduction to Timestamping: Methods, Research Process and Ethics

Zach Morrison, PhD Student, SpokenWeb Research Assistant, University of Alberta

By timestamping literary audio performances, SpokenWeb researchers transform these lengthy and often unwieldy objects into navigable digital files amenable to future critical engagement. Timestamps index the speakers, literary works and topics that appear in each sonic artifact, allowing scholars to easily locate and listen to the spans of audio relevant to their research interests. They therefore improve the accessibility of SpokenWeb's extensive archive and facilitate the activation of archival objects in the present. In this section, I will detail the various steps of the time-stamping process using a 1979 poetry performance by Fred Wah, a Canadian poet of mixed Swedish, Irish-Scots and Chinese heritage, as a case study. Drawing upon Wah's performance, I will discuss not only the methods and types of research that are necessary to appropriately label literary audio events, but also the difficulties encountered when attempting to describe spans of audio that do not neatly fit into the syntax of SpokenWeb's style guide. Crucial to this discussion will be the ethics of timestamping, as I will attend to the omissions effected by the timestamp itself, the small laughters, loaded pauses, and other sonic excrescences that exceed its limited descriptive space, and the potential

for capturing these minor events as SpokenWeb's practices continue to evolve.

5. Audio Rights Management and Ethical Curation

Michael O'Driscoll, Member, SpokenWeb Governing Board; Professor, Department of English and Film Studies, UAlberta

The SpokenWeb Project is committed to the development and dissemination of digital audio objects drawn from analogue collections of literary performance located at institutions across Canada. In almost all cases, these recordings have never been publicly available, and their migration to an, ideally, open digital environment carries an extraordinary legal and ethical weight. The Project's cross-institutional Rights Management Task Force (of which I am a member) is responsible for advising and supporting community and institutional partners in developing relationships, practices, and mechanisms that ensure open access to recordings that respect Canadian copyright law, rights holders, and performers. The legalities of rights management in the case of audiotextual performance are murky at best, and while the law seems to favour the recordist over the content creator, the ethical implications of curating such archived events stress a nonetheless heightened responsibility. The SpokenWeb collection includes everything from public events such as literary readings, classroom lectures, and panel presentations to more private events such as informal interviews, casual coffee-table recordings, and the sometimes revealing and even heated discussions that constitute the paratextual elements of exchanges captured on tape. Given the age of the recordings, the sometimes vulnerable identities of the speakers, and the ongoing interests of the (often still living) creative artists involved, curating these collections requires a delicate balance between the prerogatives of rights management and the responsibilities of ethical curation that devolves, often, to the specificity of the collections and audio objects under our stewardship.

6. Lightening the Load: Minimalist Web Design and Development for Scholars

Chelsea Miya, PhD, Postdoctoral Fellow, SpokenWeb, University of Alberta

Tejas Ambarani, MDES Student, SpokenWeb Research Assistant, University of Alberta

Knowledge dissemination, as a key component of scholarly work, increasingly occurs through online channels. When showcasing your research on the web, it is tempting to rely on tools like WordPress, Wix, and Squarespace, which promise glossy, ready-made builds at your fingertips. However, as Alex Gil points out, these user-friendly apps have "disconnected" researchers from the "material conditions of their own knowledge production" (Gil). Not only that, these tools are often server-heavy, stacking multiple programs on top of one another, and as such need to be continually updated. Drawing on the minimalist computing philosophy of Gil, Jentery Sayers, and others, we will use SpokenWeb UAlberta as a test-case for how to go "back to the basics" (Sayers) and build a website from scratch using basic programming languages. In our paper, we will explain how to implement minimal computing principles in both the front and back-end of development, creating a website that is self-sufficient, low-maintenance, and accessible. In addition to the underlying architecture, we will also describe how to approach the look and feel of a website with restraint, using less intrusive elements like flat colours and vector-based illustrations to cut out "excess" and foreground "just content" (Sayers). We will finally explain how to be more selective about features without compromising user experience and walk-through the benefits of using personas and scenarios to guide the design process.

Bibliography

Sayers, Jentery. *Minimal Definitions*. https://jntry.work/mindefinitions/.

Gil, Alex. "The User, the Learner and the Machines We Make." GO: DH. http://go-dh.github.io/mincomp/thoughts/2015/05/21/user-vs-learner/

CLIP and beyond: Multimodal and Explainable Machine Learning in the Digital Humanities

Offert, Fabian
offert@ucsb.edu
University of California, Santa Barbara, United States of America

Impett, Leonardo
leonardo.l.impett@durham.ac.uk
Durham University, United Kingdom; Cambridge University, United Kingdom

Al Moubayed, Noura
noura.al-moubayed@durham.ac.uk
Durham University, United Kingdom

Cetinic, Eva
eva.cetinic@irb.hr
Ruđer Bošković Institute, Croatia

Bell, Peter
peter.bell@uni-marburg.de
Marburg University, Germany

Smits, Thomas
thomas.smits@uantwerpen.be
University of Antwerp, Belgium

Leone, Anna
noura.al-moubayed@durham.ac.uk
Durham University, United Kingdom

Watson, Matthew
noura.al-moubayed@durham.ac.uk
Durham University, United Kingdom

Winterbottom, Tom
noura.al-moubayed@durham.ac.uk
Durham University, United Kingdom

Kluvanec, Dan
noura.al-moubayed@durham.ac.uk
Durham University, United Kingdom

Lawrence, Dan
noura.al-moubayed@durham.ac.uk
Durham University, United Kingdom

Kosti, Ronak
peter.bell@uni-marburg.de
University of Erlangen-Nuremberg, Germany

Wevers, Melvin
thomas.smits@uantwerpen.be
University of Amsterdam, the Netherlands

Lefranc, Lith
thomas.smits@uantwerpen.be
University of Antwerp, Belgium

Panel Introduction (Fabian Offert, Leonardo Impett)

Until very recently, the computational analysis of text and images have been regarded as two entirely separate areas of research within the digital humanities (DH), mirroring the technical separation of natural language processing and computer vision in computer science. In the age of deep learning, this separation has begun to erode, as models increasingly become more general (e.g. Lu et al. 2021) and more multimodal (e.g. Dosovitskiy et al. 2020).

This development towards an integration of text and images has culminated in the release of the CLIP (Contrastive Language–Image Pre-training, Radford et al. 2021) model by OpenAI at the beginning of 2021. CLIP allows us to study images in a linguistic context, and vice versa. Applications include zero-shot labeling, semantic clustering, and zero-shot object generation, among others. In conjunction with more established generative techniques like GANs (Goodfellow et al. 2014) and diffusion models (Dhariwal et al. 2021), CLIP even facilitates the prompt-guided generation of images from scratch, as evidenced by the recent emergence of CLIP-based AI artworks on the web. By generating images that seek to maximize the activation of a specific neural network, these CLIP-generated images overlap heavily with techniques from interpretable machine learning (see e.g. Molnar 2020, Doshi-Velez and Kim 2017); and the generated images themselves give scholars in the digital humanities new tools to interpret the implicit visual culture in large neural models.

Consequently, DH researchers are now beginning to integrate CLIP in particular, and multimodal models in general, into their research. Recent digital humanities projects utilize CLIP to automatically classify cultural heritage datasets with no specific training, to reimagine contemporary artworks based on their titles alone, and to explore large image corpora with natural language prompts (e.g. Offert 2021). CLIP has also significantly facilitated the development of new explainability techniques that promise to further consolidate computational (distant) and hermeneutical (close) approaches in DH (see Liu 2013).

The proposed panel reflects on this development by bringing together researchers from computer science, digital art history, computational literary studies, and related disciplines, facilitating an interdisciplinary discussion on the current state of multimodal machine learning and the potential of models like CLIP for DH. Importantly, the panel aims to not only discuss practical aspects of CLIP and related models but also to evaluate their clear limitations and inherent biases. Moreover, the panel seeks to provide a space for the discussion of the epistemological implications

of such models, focusing in particular on the increasing reliance of the digital humanities on pre-trained models and inaccessible large-scale datasets from computer science, and the "downstream" effects of this dependency. Finally, the panel proposes to examine the artistic potential of CLIP and its implications for DH, including the significant lack of a proper conceptual apparatus to evaluate projects at the intersection of the digital humanities and creative practice.

Contributions to the panel investigate these broader topics in relation to specific digital humanities projects and questions, including explainable machine learning models within archaeology, the potential for CLIP in nuancing gender classification, the role of CLIP-generated images in the digital humanities as both artworks and diagnostic tools, and CLIP as a case study for the epistemological analysis of deep learning models within the digital humanities.

Multimodal Deep Learning Meets Digital Art History (Eva Cetinic)

Multimodality is inherent to almost all aspects of human perception, communication, and production of information. However, as a phenomenon, multimodality is particularly important for the epistemological, interpretive and creative processes within art and art history. The historical beginning of multimodality research in art history can be traced back to Lessing's Treatise on Laocoön (1766) and the discussion of spatio-temporal differences of poetry and painting. Modern multimodality research emerged from the field of functional linguistics and evolved in the last three decades into established theoretical frameworks, linked to various other disciplines such as semiotics, media studies or information design. However, in the context of humanities, most theories of multimodality lack strong empirical foundations and might therefore potentially benefit from embracing computational methods for building and analysing large multimodal data collections. In the context of computer science, multimodal machine learning is a well-established field (see Baltrušaitis et al. 2018) which has very recently been revolutionized with the introduction of transformer-based large-scale vision-language pre-trained models, such as CLIP (Radford et al. 2021). This paper discusses how such models can be integrated with methodological practices in the domain of digital humanities. In particular, the paper shows how CLIP can be used to analyze complex relations between aspects of multimodal objects in digitized art collections. Furthermore, the paper aims to discuss how CLIP can be utilized to produce new digitally-born content and novel navigational mechanisms in virtual artistic spaces, with specific reference to one of its first implementations in this context, namely the "The Next Biennial Should be Curated by a Machine" project (Krysa and Impett 2021).

"CLIP Studies": Analyzing Large-scale Deep Learning Models in the Digital Humanities (Fabian Offert)

Pre-trained deep learning models have become important tools for exploratory data analysis. Replacing earlier attempts at sorting large search spaces by formal aspects like color (see Manovich 2020), neural network architectures like Inception (Szegedy et al. 2015) and VGG (Simonyan and Zisserman, 2014) have significantly improved the semantic clustering of images. OpenAI's CLIP model (Radford et al. 2021) represents another improvement over these approaches, both in terms of the quality of its image embeddings and its ability to relate image and text. Thus, CLIP promises to become a de-facto standard in digital art history (see Brey 2021, Brown 2020). At the same time, the black-box character of earlier visual models (Offert and Bell 2020) is amplified in CLIP, which has been trained on proprietary data sources and cannot be retrained on consumer hardware. Taking up this development, the paper suggests that specific models like CLIP have become "influential" enough to warrant a dedicated epistemological analysis. Echoing Alan Liu's call for a "close reading of distant reading" (Liu 2020), the paper argues that such an analysis needs to be separate from applied DH work but also cannot be "outsourced" to disciplines like media studies and science and technology studies. Concretely, such an analysis needs to reach beyond the established call for "datasheets" (Gebru et al. 2018) or "model cards" (Mitchell et al. 2019) that specify training data sources and potential biases. It needs to address the inductive biases of a model's underlying architecture and include reproducible tests (using standardized test datasets). Most importantly, it needs to make use of existing interpretability techniques, including generative approaches. Taking all this into account, the paper sketches a preliminary epistemological analysis of the CLIP model as a first case study.

Debinarizing Gender Classification: Teaching CLIP to Postpone Binarization as an Algorithmic Quality (Thomas Smits, Lith Lefranc, Melvin Wevers)

In cultural theory, scholars have conceptualized gender identities and the ways in which they find (visual) expression in heritage collections, as non-binary socio-cultural constructs (Matsuno and Budge 2017). In contrast,

most applications of machine learning in digital humanities classify gender into two mutually exclusive classes. Common performance metrics further exacerbate binarity by penalizing models for uncertainty and non-response. This paper uses CLIP (Radford et. al 2021), a multimodal model, in combination with C@1 (Peñas and Rodrigo 2011), a F1 metric that allows (and rewards) non-response, to propose a new method for gender classification on (historical) images. Binary (gender) classification is built on the conceptual fallacy that a 0.05 prediction for class A automatically entails a 0.95 prediction for class B. We previously showed that CLIP can only simulate (binary) classification tasks (Smits and Kestemont 2021). CLIP can only approach binary classification by asking two questions (Is this A? Is this B?) and normalizing the outcomes into a single prediction. By measuring CLIP's approximation of binary prediction with C@1, we hypothesize that we can calibrate algorithms to know when they do not have enough information to make a binary prediction (self-awareness), or when they should postpone binarization. We test our recalibrated algorithm on stratified sets of nineteenth-century magic lantern slides (Smits and Kestemont 2021), mid-twentieth century advertisements (Wevers and Smits 2020), and 'modern' photographs scraped from the internet (Schumann et. al 2021). We hope this helps to shed light on the ways in which gender functioned as an historical social-cultural construct.

Explainability in Deep Learning for Archaeology (Anna Leone, Noura Al Moubayed, Matthew Watson, Tom Winterbottom, Dan Kluvanec, Dan Lawrence)

The rise of explainable Machine Learning (ML) has seen the development of tools that aim to decipher decisions made by black-box ML models. These techniques can be used to both understand and verify these decisions by providing interpretable outputs from ML models. They highlight which features of the input were deemed most (and least) important by the model. Explainable ML has also been used to help better understand the limitations of these models, providing the basis for model improvement. Thus, the use of explainable ML can increase the understanding of, and trust placed in, ML models; especially in applications where ML-expertise is not expected of the end user. We discuss applying explainability to ML models trained on a number of varying archaeology-based tasks and how this could aid their wider adoption. For example, in our model for generating artefact metadata such as "this artefact is from Iraq", explanations are produced that highlight regions of interest that attempt to explain 'why' the model believes the artefact to be from Iraq. These explanations can then be compared to explanations from domain experts to confirm the model is looking at the correct parts of the image. Similar techniques could be used to identify which parts of the image were most useful when identifying (possibly) stolen artefacts, and to highlight parts of the images that were most useful when retrieving similar images. These examples showcase how important explainability is when it comes to increasing understanding, and hence trust, of ML models to non-ML experts.

Do Parrots Dream of Electric Sheep? (Leonardo Impett)

OpenAI published two new models on January 5th 2021: CLIP (Radford et al. 2021) and DALL·E (Ramesh et al. 2021). Whilst CLIP (which is open source) focuses on calculating image-text similarity, DALL·E is a closed-source pipeline (postprocessed with CLIP) for generating images based on texts. DALL·E generated at least as much public excitement as CLIP, and soon a host of community solutions, based on the public CLIP model, had been proposed to generate images: including Ryan Murdock's BigSleep (CLIP-guided BigGAN) and Aleph (which reuses the only open-source part of DALL·E, the autoencoder), Phil Wang's Deep-Daze (CLIP and SIREN), Katherine Crowson's CLIP+VQGAN and CLIP-guided image diffusion models; and a host of others in 2022 (including DALL·E 2). These image generation systems build on interpretable machine learning techniques such as DeepDream (Mordvintsev et al. 2015). A key role for CLIP-guided image models within the digital humanities is as a window onto deep neural models of contemporary visual culture - seeming to know more about Studio Ghibli than Ghiberti. This paper will argue that text-guided image generation speaks to two important debates within critical AI studies: Molyneux's problem, and the Chinese room argument (Searle 1980). The first concerns the relationship between seeing and knowing, and the role of knowledge (as probabilistic priors) in vision; and the second on the relationship between symbolic and embodied knowledge. Purely symbolic models can be dismissed as stochastic parrots; but models that can visualise, as well as describe, new situations offer us a far deeper view on the cultural and ideological assumptions of large neural networks.

AI Art and its Limits (Peter Bell, Ronak Kosti)

Pre-training on large-scale unlabeled datasets has proven quite useful recently for language (GPT-3, Brown et al. 2020), vision (ViT, Dosovitskiy et al. 2020) and language

+vision (CLIP, Radford et al. 2021) models alike. With networks like IIN (disentangling invertible interpretation network, Esser et al. 2020), it has become possible to use expert language and vision models in conjunction while increasing their interpretability. At the same time, the general public's fascination with the CLIP+VQGAN generative model, which is not unfounded, has led to a glut of "AI Art" on social media. But how can this so-called generative art be classified? Is it the artist (original source of inspiration for the AI), the composer of the text prompt, or is it the algorithm itself that determines its aesthetic status? In a series of experiments we confront CLIP+VQGAN (and other networks) with works and concepts from art history. We evaluate the data biases that may have seeped into the generative aspects of these models, using different models trained on large-scale image sets like ImageNet, COCO, Open-Images and Flickr. Furthermore, we investigate the general problem of the a-historical training of CNNs via contemporary training sets by using art historical iconographies and topics in our prompt lines. We observe that the networks are capable of blending various cultural concepts but are easily misled by polysemy and biases. Their generative aspects also suffer from mis-attribution of basic concepts, as well as prudent localization. Hence, we suggest the necessity of a "critical machine vision" approach that combines methods of interpretation from both an art historical and a technical perspective.

Bibliography

Baltrušaitis, T., Chaitanya A., and Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423-443.

Brey, A. (2021). Digital art history in 2021. *History Compass* 19(8).

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., and Askell, A. (2020). Language models are few-shot learners. arXiv preprint 2005.14165.

Brown, K., ed. (2020). *The Routledge Companion to Digital Humanities and Art History*. Routledge.

Dhariwal, P., and Nichol, A. (2021). Diffusion models beat GANs on image synthesis. arXiv preprint arXiv:2105.05233.

Doshi-Velez, F., and Kim B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint 1702.08608.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M. et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint 2010.11929.

Esser, P., Rombach, R. and Ommer, B. (2020). A disentangling invertible interpretation network for explaining latent representations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*: 9223-9232.

Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumeé, H., and Crawford, K. (2018). Datasheets for datasets. arXiv preprint 803.09010.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*: 2672-80.

Joasia, K. and Impett, L. (2021). The next Biennial should be curated by a machine - A research proposition. *Stages* 9.

Liu, A. (2020). Humans in the loop: Humanities hermeneutics and machine learning. DHd 2020 keynote. URL: https://liu.english.ucsb.edu/humans-in-the-loop-dhd2020-conference/.

Liu, A. (2013). The meaning of the digital humanities. *PMLA* 128(2): 409-23.

Lu, K., Grover, A., Abbeel, P., and Mordatch, I. (2021). Pretrained transformers as universal computation engines. arXiv preprint 2103.05247.

Manovich, L. (2020). *Cultural Analytics*. MIT Press.

Matsuno, E., and Budge, S. L. (2017). Non-binary/genderqueer identities: A critical review of the literature. *Current Sexual Health Reports* 9(3): 116-120.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*: 220-229.

Molnar, C., Casalicchio, G. and Bischl, B. (2020). Interpretable machine learning - A brief history, state-of-the-art and challenges. arXiv preprint 2010.09337.

Offert, F. (2021). imgs.ai - a fast, dataset-agnostic, deep visual search engine for digital art history. URL: https://imgs.ai.

Offert, F., and Bell, P. (2020). Perceptual bias and technical metapictures. Critical machine vision as a humanities challenge. AI & Society 36: 1133-1144. URL: https://link.springer.com/article/10.1007/s00146-020-01058-z

Peñas, A., and Rodrigo, A. (2011). A simple measure to assess non-response. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*: 1415-1424.

Radford, A., Kim J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G. et al. (2021). Learning transferable visual models from natural language supervision. arXiv preprint 2103.00020.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021). Zero-shot text-to-image generation. arXiv preprint 2102.12092.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences* 3(3): 417-24.

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint 1409.1556.

Smits, T., and Kestemont, M. (2021). Towards multimodal computational humanities. Using CLIP to analyze late-nineteenth century magic lantern slides. *Proceedings of CHR 2021: Computational Humanities Research Conference, November 17–19, 2021, Amsterdam, The Netherlands*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich. A. (2015). Going deeper with convolutions. *Computer Vision and Pattern Recognition (CVPR)*.

Wevers, M., and Smits, T. (2020). The visual digital turn: Using neural networks to study historical images. *Digital Scholarship in the Humanities* 35(1): 194-207.

Books' Impact in Digital Social Reading: Towards a Conceptual and Methodological Framework

Pianzola, Federico
f.pianzola@rug.nl
University of Groningen, Netherlands, The

Viviani, Marco
marco.vivani@unimib.it
University of Milano-Bicocca, Italy

Fossati, Alessandro
a.fossati@campus.unimib.it
University of Milano-Bicocca, Italy

Boot, Peter
peter.boot@huygens.knaw.nl
Huygens Institute for the History of the Netherlands, Netherlands, The

Fialho, Olivia
olivia.fialho@huygens.knaw.nl
Huygens Institute for the History of the Netherlands, Netherlands, The; Utrecht University, Netherlands, The

Koolen, Marijn
marijn.koolen@gmail.com
Huygens Institute for the History of the Netherlands, Netherlands, The; KNAW Humanities Cluster, Netherlands, The

Neugarten, Julia
j.neugarten@gmail.com
Huygens Institute for the History of the Netherlands, Netherlands, The

Van Hage, Willem Robert
w.vanhage@esciencecenter.nl
Netherlands eScience Center, Netherlands, The

Rebora, Simone
simone.rebora81@gmail.com
Università degli Studi di Verona. Italy

Herrmann, J. Berenike
berenike.herrmann@uni-bielefeld.de
University of Basel, Switzerland

Messerli, Thomas C.
thomas.messerli@unibas.ch
University of Basel, Switzerland

Jorschick, Annett
annett.jorschick@uni-bielefeld.de
University of Basel, Switzerland

Sharma, Srishti
srishti.0118@gmail.com
Independent scholar

The aim of this panel is to debate the challenges and opportunities offered by online reviews for measuring the impact that books can have on readers (Boot and Koolen, 2020). The focus is specifically on culture- and language-specificity, thus we will compare insights from the analysis of Korean, English, Italian, German, and Dutch reviews.

Digital social reading platforms – like Goodreads, Lovelybooks, or Naver Books – host millions of reviews and, thus, offer unique possibilities for research into literature, reading, and reader response (Rebora et al., 2021; Walsh and Antoniak, 2021). Computational tools are especially relevant, given the large amount of available data, but finding associations between textual features, cultural conventions (e.g. genre), and cognitive, affective,

and aesthetic responses is not a straightforward task (Koolen et al., 2020; Pianzola et al., 2020).

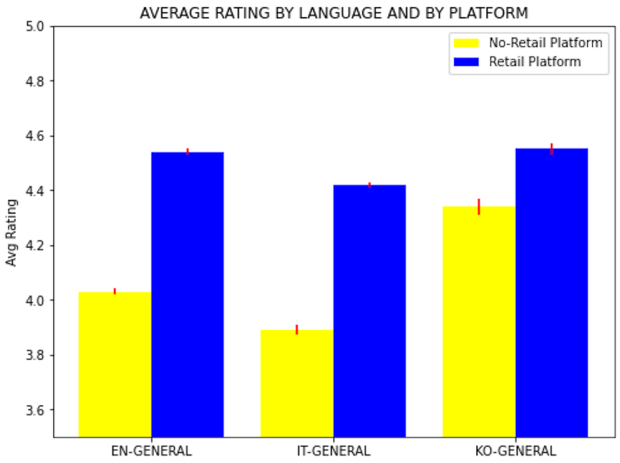
By comparing research done with different platforms, datasets, and languages, we aim at improving the methods that we employ, in a dialogue involving both data-driven insight and theoretical reflection on literature and readers. Questions that we will address are: what aspects of a book's impact on readers can reviews help us to measure? What are the limitations of online book reviews for studying impact? How do we know to what extent these review texts reflect the actual reading experiences? What are unwanted, confounding influences (e.g. reviewers projecting a favourable self-image, socially desired responses, aspects of identity formation, fake reviews). How do online book reviews differ from experimentally controlled gathering of reader responses (lab studies, questionnaires, psychologically validated scales) (Lendvai et al., 2020)? How do platforms for reviewing and social interactions around books influence reviewers and their perceptions? How do reviewers compare to other readers?

To answer such questions, we will present four case studies dealing with different languages and cultures, followed by an open discussion of the results and methods, reflecting on their generalizability, efficacy, and limitations.

Cross-cultural and Multilingual Book Reading and Reviewing: Building and Analyzing a Dataset for English, Italian, and Korean

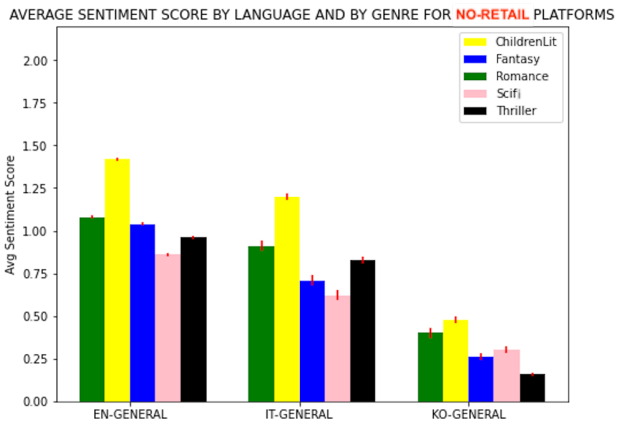
Fossati, A., Pianzola, F., Viviani, M.

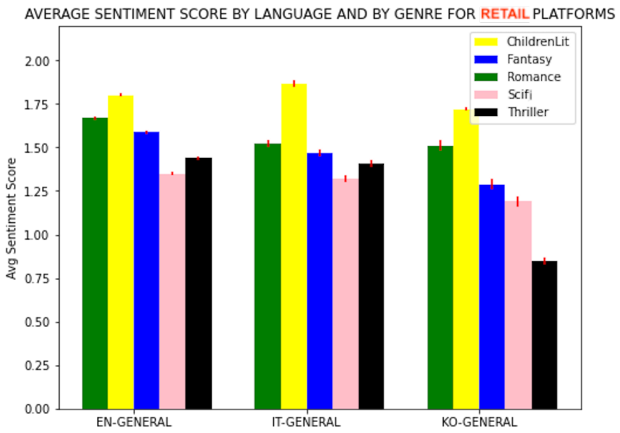
In this project, we analyze the differences between English, Italian, and Korean speaking readers in relation to the impact that a book has on them, namely in relation to the attitude that reviewers of different cultures have in providing their opinion about books on a digital platform. To this aim, we scraped reviews from the biggest reviewing platforms for each language: Amazon.com (270k reviews collected) and Goodreads (247k) for English (999 books collected), Amazon.it (93k) and Anobii (64k) for Italian (975 books), and Naver Books (40k) and Yes24 (67k) for Korean (900 books). We sampled one retail and one non-retail platform for each language because one of our goals was to reproduce the results by (Dimitrov et al., 2015; Newell et al., 2016) about the differences in readers' behavior (Fig. 1).



Average rating from 3,000 books in English, Italian, and Korean taken from retail (blue) and no-retail (yellow) platforms. Reproduction of the results obtained for English by (Newell et al. 2016).

We are providing an important cross-cultural and multilingual resource and analytical contribution to this kind of research. We present the process of construction and preliminary analyses of this multilingual corpus, which is aimed both at containing common books (about a thousand for each language, with their respective reviews) in all three languages, and at highlighting reading preferences in terms of genres (books from Children, Romance, Sci-fi, Thriller and Fantasy genres) and authors that are peculiar for each language/culture (Fig. 2). This kind of dataset is necessary – but so far unavailable – to implement analyses that could reliably explore the impact that books have on both Western and Asian readers.





Average sentiment scores for books belonging to 5 different genres. Values are computed using a transformer-based multilingual model (XLM-R) specifically fine-tuned for sentiment analysis of book reviews.

Reading Impact in Online Book Reviews: Challenges and Prospects

Boot, P., Fialho, O., Koolen, M., Neugarten, J., Van Hage, W.R.

What is the impact of fiction? In the *Impact and Fiction* project we investigate that question by measuring impact in a corpus of 500k+ online book reviews and relating this to high-level features (mood, topic, style, narrative) computationally extracted from a corpus of novels discussed in these reviews. In predicting the impact we also take reader features into account. We draw from a growing tradition of studies on the impact of reading fiction, including the phenomenological and experimental tradition (e.g. Miall and Kuiken, 1995; Kuijpers, 2014; Fialho, 2012, 2019), studies of literary evaluation (e.g., Von Heydebrand and Winko, 1996), of newspaper criticism (Linders, 2012), of online reviews, depending on site and book genre (Koolen et al., 2020; Newell et al., 2016), of the influence of reader gender on reviews (Thelwall and Boerrier, 2019), and of self-presentation in social media (e.g. Hollenbaugh, 2021).

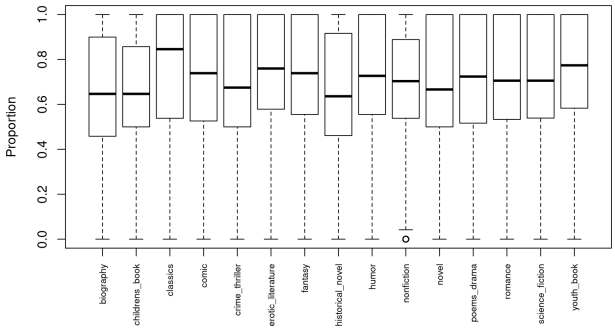
Despite the wealth of research in these domains, there are many open questions on the impact of fiction in the context of book reviews. Are existing typologies of reading experiences applicable to the context of book reviews? Can impact be reliably measured from reviews? How do reviewer characteristics and textual features (e.g., genre, perspective) affect impact? Do genre effects influence review content? Are book reviews mostly about books, or do they primarily reflect the self-image readers want to present to the world? What forms of reflection occur in book reviews? In this presentation, we will offer a series of reflections on these issues.

Dealing with messy data. A methodological solution for analysing unbalanced social media datasets

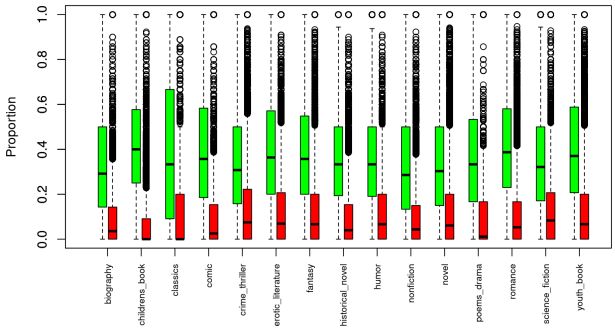
Rebora, S., Hermann, J. B., Messerli, T., Jorschick, A.

In the study of natural social media data standard solutions for dealing with "messy" and high frequency data in hypothesis-testing statistics are still missing. This paper contributes to a solution for the issues of hypothesis testing of (a) big-scale and (b) unbalanced datasets. Building on the development of deep learning classifiers for the recognition of evaluative language and sentiment (Rebora et al., 2022), we thus present a possible methodological groundwork for the study of book impact at a large scale.

Our project focuses on German book reviews published on the *LovelyBooks* platform (~1.3M reviews). Fig. 3 (evaluation) and 4 (sentiment) show the application of the two classifiers on the corpus.



Proportion of evaluative language per review

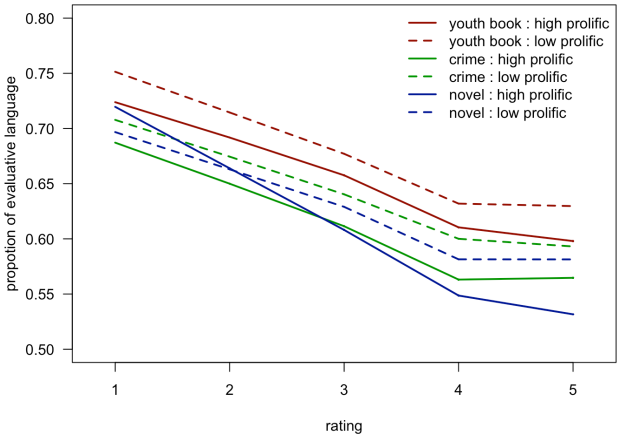


Proportion of sentiment per review (green = positive; red = negative)

In order to deal with the issue of a "messy" big-data corpus, we evaluated advanced statistical strategies. As Anova-style tests tend to increase type I errors, we applied Linear Mixed-Effects Models (Winter and Grice, 2019), using a subsection of our dataset (30,000 reviews, most popular genres). Book GENRE, RATING, and total Number of Reviews by User (NRU) were independent variables predicting the proportion of evaluative language. Table 1 shows significant main effects for *NRU, RATING, GENRE*, and a significant interaction effect for *NRU* and *GENRE*. Figure 6 shows an interaction effect of GENRE, RATING and USERTYPE (high vs. low NRU), with high NRU users deviating for the novel GENRE. As such dynamics might often be missed in data analysis, our case study shall advocate for the use of advanced statistical modeling.

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
NRU	8.4	8.4	1	15444.09	15.78	**<.001**
RATING	329.1	82.3	4	28455.70	153.70	**<.001**
GENRE	6.6	3.3	2	25034.65	6.15	**.002**
NRU : RATING	0.4	0.1	4	29220.88	0.20	.937
NRU : GENRE	3.8	1.9	2	29064.25	3.54	**.029**
RATING : GENRE	5.4	0.7	8	28245.37	1.27	.253
NRU : RATING : GENRE	7.8	1.0	8	29403.24	1.82	.068

Anova type III table of main effects and interactions. Significant effects are bold.



Rating vs. evaluative language

A preregistered analysis investigating the relation between emotions in books and reviews

Sharma, S., Pianzola, F.

In this paper we look at the emotional impact a book can have, focusing on the relation between emotions in books and emotions in reviews. We test the psychological theory known as "framing effect" (Tversky and Kahneman, 1981) – which states that the response of the audience is influenced by the way in which the information is presented to them – and operationalize it as a task of computational text analysis. To do this, we use a dataset of 450 books, divided across 9 genres, and with more than 5 million English reviews. We conduct sentiment analysis of three different components: the average book sentiment, the average review sentiment of the corresponding book, and the emotional story arc of each book (Reagan et al., 2016; Jockers, 2017). We compare three different methods – distilBERT (Sanh et al., 2020), a dictionary-based model testing different lexica (Mohammad and Turney, 2013), and SentiArt, a vector space model (Jacobs and Kinder, 2019) – and reflect on their accuracy and interpretability in the context of a DH project, rather than as a general NLP task. This is among the first studies that quantitatively investigate relations between stories' sentiment and reader response (Jacobs and Kinder, 2019; Pianzola et al., 2020), and it uses both state-of-the-art machine learning as well as hypothesis-testing statistics.

Bibliography

Dimitrov, S. et al. (2015). Goodreads versus Amazon: The Effect of Decoupling Book Reviewing and Book Selling. *Proceedings of the International AAAI Conference on Web and Social Media*, **9**(1), 602-05. https://ojs.aaai.org/index.php/ICWSM/article/view/14662

Fialho, O. (2012). Self-Modifying Experiences in Literary Reading: A Model for Reader Response. PhD Dissertation, University of Alberta. https://era.library.ualberta.ca/items/94ecceb5-56a7-4601-b1e5-bc2117d12e01

Fialho, O. (2019). What is literature for? The role of transformative reading. *Cogent Arts & Humanities,* **6**(1), special issue "The place of the cognitive in literary studies", Kukkonen, K., Kuzmičová, A., Ledet Christiansen, S., and Polvinen, M. (eds.), https://doi.org/10.1080/23311983.2019.1692532

Hollenbaugh, E. E. (2021). Self-Presentation in Social Media: Review and Research Opportunities. *Review of Communication Research*, **9**, pp. 80-98.

Jacobs, A. M. and Kinder, A. (2019). Computing the Affective-Aesthetic Potential of Literary Texts. *AI*, **1**(1), pp. 11–27. 10.3390/ai1010002.

Jockers, M. (2017). *Introduction to the Syuzhet Package*. *The Comprehensive R Archive Network*. https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html (accessed 29 May 2018).

Koolen, M., Boot, P., and Van Zundert, J. (2020). Online Book Reviews and the Computational Modelling of

Reading Impact, *Computational Humanities Research 2020*. http://ceur-ws.org/Vol-2723/long13.pdf

Kuijpers, M. (2014). Absorbing stories: The effects of textual devices on absorption and evaluative responses. PhD Dissertation, Utrecht University.

Linders, Y. (2012). Argumentation in Dutch literary criticism 1945–2005. In C. Perry and M. Szurawitzki (eds.), *Sprache und Kultur im Spiegel der Rezension* (Frankfurt am M.: Peter Lang), pp. 261-68.

Miall, D. S. and Kuiken, D. (1995). Aspects of literary response: A new questionnaire', *Research in the Teaching of English*, pp. 37-58.

Mohammad, S. and Turney, P. (2013). Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, **29**(3), pp. 436–65. http://dx.doi.org/10.1111/j.1467-8640.2012.00460.x.

Newell, E. et al. (2016). To Buy or to Read: How a Platform Shapes Reviewing Behavior. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)*, pp. 643–6.

Pianzola, F., Rebora, S., and Lauer, G. (2020). Wattpad as a Resource for Literary Studies. Quantitative and Qualitative Examples of the Importance of Digital Social Reading and Readers' Comments in the Margins. *PLoS ONE*, **15**(1). https://doi.org/10.1371/journal. pone.0226708.

Rebora, S., Messerli, T. C. and Herrmann, J. B. (2022). Towards a Computational Study of German Book Reviews. A Comparison between Emotion Dictionaries and Transfer Learning in Sentiment Analysis. DHd 2022 Kulturen des digitalen Gedächtnisses. 8. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum" (DHd 2022), Potsdam. https://doi.org/10.5281/zenodo.6328141

Reagan, A. J. et al. (2016). The Emotional Arcs of Stories Are Dominated by Six Basic Shapes. *EPJ Data Sci.*, **5**, pp. 5–31. 10.1140/epjds/s13688-016-0093-1.

Sanh, V. et al. (2020). DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *ArXiv:1910.01108 [Cs]*. http://arxiv.org/abs/1910.01108 (accessed 1 June 2020).

Thelwall, Mike and Bourrier, Karen (2019), The reading background of Goodreads book club members: a female fiction canon? *Journal of Documentation*, **75**(5), pp. 1139-61.

Tversky, A. and Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science*, **211**(4481), pp. 453–8. 10.1126/science.7455683.

Von Heydebrand, R. and Winko, S. (1996). *Einführung in die Wertung von Literatur: Systematik, Geschichte, Legitimation* (Paderborn: Schöningh).

Winter, B. and Grice, M. (2019). Independence and generalizability in linguistics. OSF https://osf.io/zdrpc/ (accessed 24 September 2021).

Dynamics of Culture: Tracing Discourse using Computational Methods

Quinn, William Reed
quinn.wi@northeastern.edu
Northeastern University, United States of America

Messina, Cara Marta
cmessina@jsu.edu
Jacksonville State University, United States of America

Connell, Sarah
sa.connell@northeastern.edu
Northeastern University, United States of America

Blankenship, Avery
blankenship.a@northeastern.edu
Northeastern University, United States of America

Panel Overview

The promise of computational methods is to add new specificity to cultural phenomena. Machine learning, network analysis, and a variety of other approaches bring previously obscured details to light, allowing us to better understand the deep structures that shape literary production, reception, and community. The papers of this panel build off current scholarship to explore the diverse dynamics of culture: circulation, dialectics, intertextuality, and uptake.

In "Taste in Circulation: Locating Recipes in Nineteenth-Century Newspapers," Avery Blankenship shares her most recent work identifying recipe reprinting in nineteenth-century newspapers using reprinting data developed by the Viral Texts project as well as a corpus of nineteenth-century recipes pulled from printed book recipe books. Using Doc2Vec and topic tagging, Blankenship demonstrates the multiple ways in which nineteenth-century newspaper contributors use the recipe form—for culinary purposes and otherwise.

In "'Like a Starry Network': Intertextuality in Early Women's Writing," Sarah Connell shares the outcomes of a recent Women Writers Project effort to link the textual references within the Women Writers Online collection to a bibliography that currently contains more than 3,500 texts. Using drama as a case study, Connell illustrates some

insights into early women's engagements with literature culture that can be revealed by the WWP's new encoding.

In "Excluding Crossovers: Loving Blackness in Black Panther Fandom," Cara Marta Messina traces representations of race, sexuality, and gender in *Black Panther* fanfiction both including and excluding crossovers using word embedding models and network analysis. Fans who write *Black Panther* fanfiction excluding Marvel crossovers focus more on loving blackness and critiquing colonialism in a predominantly white-centered Marvel canon.

"Language as Act; Language as History" (William Reed Quinn) uses the syntactic parsing of keywords, their distances within word embeddings, and network analysis to examine the mobility of language within serialized print culture. This hybrid methodology builds off previous scholarship to better understand the semantic evolution of keywords over time by mapping them into vector space and then ascertaining their relationship to other texts with the same keyword.

Although each paper examines a different corpus, they are brought together by their shared and complementary focus on the circuits and threads that compose historical and contemporary print culture.

Contribution 1: "Taste in Circulation: Locating Recipes in Nineteenth-Century Newspapers" by Avery Blankenship

Much like today, nineteenth-century recipes tended to circulate across print media—appearing in formal books, newspapers, pamphlets and more. While the study of recipes has been well established in scholarship, many of these studies focus on formal, printed cookbooks rather than recipes that circulated in more ephemeral print publications despite their ubiquity in the nineteenth-century (see, for example, Bower 1997, Theophano 2016, Elias 2017, and Walden 2018). Particularly due to the way nineteenth-century newspapers often only roughly delineate articles, automating the detection of recipes can pose complex challenges.

Using the newspaper corpus developed by the Viral Texts project which addresses the practice of reprinting in the nineteenth-century press broadly (Smith et al. 2015), this paper uses a combination of topic tagging and Doc2Vec in order to automate the detection of recipes in nineteenth-century newspapers. Beginning with a corpus of tagged recipes, the corpus was added to a set of newspaper articles which have been hand tagged according to their genres by the *Viral Texts* team. With this training set, a Doc2Vec model was trained and used to rank unseen articles along these genres according to their similarity. Recipes identified with high similarity scores were then added to the training set so that the model could be retrained.

The many edge cases where a text is similar in form to a recipe but is not necessarily culinary reveals the complexities of this genre and its proliferation in nineteenth-century newspapers, as well as how genres "make legible the significance of particular forms" (King 2021). The goal of this preliminary analysis is to propose a more expansive definition of "recipe" to include what could be considered "recipe-adjacent" texts. These recipe-adjacent texts reveal the usefulness of "recipe" as a genre for conveying jokes, advice, agricultural information, and even storytelling. This paper will demonstrate case studies of various recipe-adjacent texts, the ways in which the model classifies them, as well as how this classification might help illuminate the complexities of this genre.

Contribution 2: "'Like a Starry Network': Intertextuality in Early Women's Writing" by Sarah Connell

In 2016, the Women Writers Project began adding bibliographic information for the thousands of textual references in the Women Writers Online collection of early women's writing. The "Intertextual Networks" project now includes markup linking all of the quotations, named titles, and citations in WWO to a bibliography of more than 3,500 texts. The project also includes expanded markup for more nuanced forms of intertextuality, such as parody, remix, and paraphrase.

The goal of this initiative is to enable research into the multivalent ways that women engaged with literate culture during the watershed period covered by WWO in which women's participation in the authorship and consumption of texts expanded dramatically (see Brayman Hackel and Kelly). We also seek to explore and theorize the representation of intertextuality through text encoding. This paper will share some initial research made possible with the 12,000 quotations, 5,900 titles, and 5,000 biblical citations in WWO, focusing on drama as a case study. The densely informational encoding for drama makes it possible to explore a range of questions. For example: do female-coded characters reference women writers more often than male characters? (they do); what genres of texts are included in dramatic speeches or stage directions? (poetry predominates, but classical texts and other dramas are well represented); and, what kinds of intertextual gestures appear in drama? (quotations are more than twice as common as named titles; biblical citations are almost nonexistent). This paper will conclude with an invitation for further

exploration and some thoughts on the research that might be pursued with this data, aiming to serve as a launching point for further work on the rich domain of early women's intertextuality.

Contribution 3: "Excluding Crossovers: Loving Blackness in the Black Panther Fandom" by Cara Marta Messina

On *Archive of Our Own* (AO3), one of the most popular websites for fans to publish their fanfiction, there are 427,391 works written in the Marvel Cinematic Universe; 95,457 works for *Captain America*; 48,793 works for *Iron Man*; and only 4,338 for *Black Panther*. In total, *Black Panther* as a fandom tag only appears in 1.03% of the Marvel Cinematic Universe (MCU) fandoms. In one of the largest fandoms on AO3 — the MCU — Black characters and stories are still overlooked. Fanfiction and AO3 often perpetuate anti-blackness in which characters are ignored, how characters are written, and which (relation-)ships are heralded. Fan studies scholars and authors have both been critiquing this pattern as well as analyzing predominantly Black fandoms to better understand how Black fan creators simultaneously critique and love mainstream texts that often do not love them back (Florini, 2019; Messina, 2021; Stitch, 2021; Thomas, 2019a; Thomas, 2019b; Wanzo, 2015).

Messina collected the 4,338 *Black Panther* fanfictions published on AO3 and split the corpus into two: *Black Panther* fanfics that included crossovers, meaning that they may appear with other MCU fandom tags and beyond, and *Black Panther* fanfics that excluded crossovers, meaning the fanfictions were written only about *Black Panther*. On the crossover corpus, Messina counts how often *Black Panther* character names appear in fanfiction, showing T'Challa and the characters in *Black Panther* are rarely central to the plot lines when both *Black Panther* and other MCU fandom tags appear. In the corpus excluding crossovers, word embedding models reveal how fans have deeper understandings of representations of Black sexualities and genders across the Black diaspora as well as how fans critique colonialism and white supremacy. Using word embedding models with words around gender, sexuality, and bodies, Messina examines the relationship between gender, sexuality, and race in ships like T'Challa/Killmonger and characters like Captain Okoye. The corpus excluding crossover demonstrates how fans carve out communities that resist white supremacy and, as bell hooks (1992) advocates for, love blackness.

Contribution 4: "Language as Act; Language as History" by William Reed Quinn

Literary theory has long held that concepts, like gender, are constructed through language and discourse. Moments of cultural divergence, then, should also be evident in language. While cultural analytics has provided many insights about cultural transitions, often across genres and generations, there remains more work to do in clarifying the details of these shifts at a finer grain.

Quinn examines divergence and semantic mobility using examples from a series of radical feminist magazines from the Modernist Journals Project. Through her editorial hand, Dora Marsden sought to change connotations around gender that went beyond the Suffragist movement. Marsden and the other editors sought to expand "feminism's scope" beyond "the vote" and to consider "gender relations, sexual life, and more broadly, women's physical, spiritual, and intellectual experience of modernity" (McMahon and Green). Marsden and the other contributors were provocateurs with a specific goal: to construct new identity categories and revise old ones.

To illustrate the editors' and contributors' attempts to reconstruct gender through discourse, Quinn finds gendered keywords (i.e., "woman," "man," "freewoman," "spinster," etc.) and inputs them with their syntactic dependencies as word embeddings. The resulting vectors and their cosine similarities can then be measured within a network to illustrate the connotative changes of key concepts over time. This experimental method moves text analysis beyond the level of document and genre in order to explore the possibilities and limits of tracking discursive evolution with computational methods.

Bibliography

Bower, Anne. *Recipes for Reading: Community Cookbooks, Stories, Histories*. Univ of Massachusetts Press, 1997.

Brayman Hackel, Heidi and Catherine E. Kelly. "Introduction." *Reading Women: Literacy, Authorship, and Culture in the Atlantic World, 1500-1800*. Ed. Heidi Brayman Hackel and Catherine E. Kelly. Philadelphia: University of Pennsylvania Press, 2008. 1–10.

Carter, Sarah. *Early Modern Intertextuality.* Cham, Switzerland: Springer, 2021.

Duhaime, Douglas. "Textual Reuse in the Eighteenth Century: Mining Eliza Haywood's

Quotations." *Digital Humanities Quarterly*. 10.1. 2016. http://www.digitalhumanities.org/dhq/vol/10/1/000229/000229.html.

Elias, Megan J. *Food on the Page: Cookbooks and American Culture*. University of Pennsylvania Press, 2017.

Florini, S. (2019). Enclaving and cultural resonance in Black "Game of Thrones" fandom. *Transformative Works and Cultures, 29*.

Green, Barbara, "The New Woman's Appetite for 'Riotous Living': Rebecca West, Modernist Feminism, and the Everyday," *Women's Experience of Modernity: 1875–1945*, ed. Ann Ardis and Leslie W. Lewis, Chicago: University of Chicago Press, 1985.

hooks, bell. (1992). *Black looks: Race and representation*. Routledge.

King, Rachael Scarborough. "The Scale of Genre." *New Literary History*, vol. 52, no. 2, 2021, pp. 261–84. *Project MUSE*, https://doi.org/10.1353/nlh.2021.0012.

McMahon, Shannon, "Freespinsters and Bondspinsters: Negotiating Identity Categories in the Freewoman," *The Journal of Modern Periodical Studies*, 2015, Vol. 6, No. 1, 2015, pp. 60–79.

Messina, C. M. (2021). The critical fan toolkit: Fanfiction genres, pedagogies, and ideologies [Doctoral Dissertation, Northeastern University]. Retrieved December 9, 2021 from http://criticalfantoolkit.org.

Porter, Dahlia. "From Nosegay to Specimen Cabinet: Charlotte Smith and The Labour of Collecting." *Charlotte Smith in British Romanticism*. Ed. Jacqueline Labbe. London: Routledge, 2015.

Thematic issue: "Digital Methods for Intertextuality Studies. *It – Information Technology*. Ed. Paul Molitor, Jörg Ritter, Stefan Conrad. 62.2. 2019. https://www.degruyter.com/document/doi/10.1515/itit-2020-0006/html

Smith, David A., et al. "Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers." *American Literary History*, vol. 27, no. 3, Sept. 2015, pp. E1–15. *Silverchair*, https://doi.org/10.1093/alh/ajv029.

Stitch. (2021 January 28). Who actually gets to "escape" in fandom?. *Teen Vogue*. Retrieved December 9, 2021, from https://www.teenvogue.com/story/who-actually-gets-to-escape-into-fandom-column-fan-service

Theophano, Janet. *Eat My Words: Reading Women's Lives Through the Cookbooks They Wrote*. St. Martin's Publishing Group, 2016.

Thomas, E. E. (2019a). *The dark fantastic: Race and the imagination from Harry Potter to the hunger games*. New York University Press.

Thomas, E. E. (2019b, May 17). Missandei, Too, Deserves Her Song – A Dark Fantastic Lament [Academic blog]. The Dark Fantastic. Retrieved December 10, 2021 fro http://thedarkfantastic.blogspot.com/2019/05/missandei-too-deserves-her-song-dark.html

Walden, Sarah W. *Tasteful Domesticity: Women's Rhetoric and the American Cookbook, 1790-1940*. University of Pittsburgh Press, 2018.

Wanzo, R. (2015). African American acafandom and other strangers: New genealogies of fan studies. *Transformative Works and Cultures, 20*.

The (Im)Possibilities of Multilingual DH in Theory and Practice: Translation, Metadata, Pedagogy

Raynor, Cecily
cecily.raynor@mcgill.ca
McGill University

Ponce de la Vega, Lidia
lidia.poncedelavega@mail.mcgill.ca
McGill University

Guénette, Marie-France
marie-france.guenette@lli.ulaval.ca
Université Laval

Kim, Eric
kimer@stanford.edu
Stanford University

Brata Roy, Samya
samyabrataroy@gmail.com
IIT Jodhpur

Dombrowski, Quinn
qad@stanford.edu
Stanford University

Introduction

The Digital Humanities offer immense possibilities for interdisciplinarity, cross-cultural, and epistemic knowledge exchange. Nevertheless, the DH continue to privilege centralized Anglophone practices and epistemologies that hinder access and contributions from marginalized communities and researchers across the globe. This panel challenges such privileges by engaging varied issues around the (im)possibilities of multilingual DH practices. The presentations in this panel contrast the multilingual aspirations of DH with the Anglophone and Anglocentric realities of DH practices regarding access, audience distributions, curation, metadata practices, pedagogical approaches, and epistemic production. Drawing upon myriad disciplinary homes, presenters in this panel engage with topics of translation, particularly translation difficulties and so-called untranslatables (Apter, 2014). For example, one panelist examines challenges in new media studies for translating contemporary DH lexica, while another maps untranslatables through computational exploration. How can languages other than English be implemented in DH? What are the barriers to entry for those engaging with non-Anglophone community translation practices? How can this linguistic negotiation be undertaken? What is gained and lost in the translation of core DH concepts? Furthermore, this panel scrutinizes the implementation gap between English and non-English DH in settings related to pedagogy and archival and metadata practices, focusing on and emphasizing the consequences of Anglocentric approaches and the advantages of multilingual DH in providing equitable access to diverse audiences, students, researchers, and linguistic communities. What does it mean to be global? How open and accessible is open access? Who can and cannot be a digital humanist when we favour English at training sites and elsewhere? How can DH transcend Anglocentrism? How can we implement multilingual DH in pedagogy and archival practices? In a range of explorations, this panel probes the possibilities and obstacles faced by multilingualism vis-à-vis English as lingua franca. In doing so, presenters in this panel engage with questions of inclusivity through a critical approach to various facets of multilingual DH.

Multilingual bio-diversity and the decolonization of the *Biodiversity Heritage Library*: The case of Latin America

Author: Lidia Ponce de la Vega, McGill University

The *Biodiversity Heritage Library* (BHL) is an online repository for global biodiversity-related literature that advocates for multilingualism but operates within (digital) Anglocentrism. With over 80% of its collection in English, the BHL promotes this language as the lingua franca of the Internet and of biodiversity-related knowledge production, positing the Global North as the epistemic center of such production and perpetuating colonial dynamics that hinder bio-diverse epistemologies from the Global South.

This presentation constitutes a reflection around best decolonial practices for online libraries and archives affiliated to the Global North but that incorporate epistemologies pertaining to the Global South. By focusing on the collections of the BHL and the case of Latin America (and Mexico, specifically), this paper discusses the role of language representation and inclusion in the decolonization of biodiversity-related knowledge production. In so doing, it explores the importance of multilingualism for the BHL in terms of access, audience distribution, curation, and epistemic production (Chan). Additionally, this presentation considers the case of BHL México, a partnership between the BHL and Mexico's Comisión Nacional para el Conocimiento y Uso de la Biodiversidad, to showcase the possibilities and limitations for the decolonization of bio-diverse online collections vis-á-vis Latin American audiences and human and nonhuman subjects–that is, our relationships with and within biodiversity. This presentation thus argues for a non-Anglocentric, non-Global-North-centric, and sympoietic (Haraway) approach to archival practices—anchored in non-hegemonic multilingualism—that aims for the diversification of bio-diverse narratives, breaks with the colonial roots of (digital) archives (Risam), and promotes a cycle of bio-diverse knowledge production in and from the Global South.

The Function of Translation in DH: Reflections from a Bangla Translator

Author: Samya Brata Roy

While engaging with the act of social media translation in DH, I found myself asking, who am I translating for? And what is the purpose of this act? Ideally, translation or an attempt at multilingualism is meant to increase reach beyond the hegemonic claws of English. But, when it comes to expressing the key terms anchored in English terminology, there often is no suitable alternative in "minor" languages, or at least, recognizable terms for non-specialist readers. Official terminologies can be valid in the literal sense, but in practical terms they can confuse readers even further. This is why using English words offers an easy way out. However, that means harking back to the dominant language. The question that comes here is: Why and how should we

translate DH, a field which exists in the disciplinary intersections? Is it counterproductive to attempt a purist translation or should one use both languages wherever necessary, a decision that presupposes an understanding of English? Sharing my experience as a Bangla Twitter translator for DHSI 21, I will raise these issues to start a conversation regarding the tussle between access (provided by a multilingual approach) and convenience (using English terminology to avoid confusion). I will build on Ortega's (2014) reflections on translating at the DH2014 conference, and Renée Desjardins' (2017) monograph on *Translation and Social Media* to ground my own ethnographic case study of social media translation. The insights drawn from my personal experience will contribute to shedding light on DH translation with regards to the positionality of the translator, in particular within minority language settings .

Pogrom: Translational and Translocational Journey of a Slavic Word through Mass Media

Author: Eric Kim, Stanford University

As Steven Zipperstein explains in the first chapter to his monograph on the Kishinev pogrom, the word *pogrom* first enters English-language mass media during the early 20th century, often accompanied by an explanatory note or offset with italics to indicate the foreignness of the term. With the growing migration of Jews beginning in the 1880s, the word eventually developed a currency and began to signify antisemitic violence on its own (Zipperstein). However, *pogrom* in the Russian context has never conveyed this specific meaning, but rather, it has referred, and continues to refer, to any instance of government-organized mass violence against groups determined along class, social, or other boundaries (Ushakov). In order to specify the Jewish victims of the riot, Russian newspapers necessarily attached the adjective Jewish, *evreiskii*, to the word *pogrom*.

By comparing the distributions and appearances of *pogrom* in variously languaged print media, I hope to explore the different perceptions of the word and the differing instantiations of antisemitic violence across national borders. Included in this analysis will be other terms synonymous with *pogrom* to signify antisemitic catastrophe, such as *besporiadok* in Russian or *riot* in English, and use of images to depict these traumatic events, while the initial corpora of interest are *Ogonek* and *Life*. Between the distinct uses and frequencies of the term in these two periodicals, I aim to find patterns in subject formation, both of the self and the other, through levied accusations, and also glaring unacknowledgments, of antisemitism.

Strategies for Multilingual DH Pedagogy

Author: Quinn Dombrowski, Stanford University

The DH 2019 pre-conference workshop on pedagogy highlighted efforts across the globe to put digital tools and methods into the hands of students. While "international" pedagogical spaces (such as summer or winter workshops, or open-access course materials) still most commonly treat English as the default language of instruction and most likely object of study, efforts to cultivate learning communities centered on materials in other language have recently expanded, including *Digital Humanities for Japanese Culture* at DHSI in 2019 (Kiyonori Nagasaki et al.), *East Asian Studies and Digital Humanities* at DReAM Lab in 2021 and 2022 (Paula R. Curtis and Paul Vierthaler), and Slavic DH workshops at the Herder Institute in Marburg, Germany and Princeton University in 2018 and 2019 (Natasha Ermolaev et al.) These linguistically-focused pedagogical encounters are invaluable for both skill development and community-building, but they are limited in each: far more students are exposed to DH through general-purpose "intro to DH" courses that do not center language-specific issues. As a result, students with research materials in English (or the dominant language of the course) leave with practical skills, while students working with other languages face an implementation gap when they try to apply what they've learned. This talk will draw on experiences teaching a deliberately multilingual DH course and participating in language-centric workshops, in order to propose actionable strategies for making DH survey courses better support the kinds of multilingual work described in the other talks on this panel.

Multilingual DH as a Political, Cultural and Ideological Statement on Accessibility

Authors: Marie-France Guénette, Université Laval; Cecily Raynor, McGill University

What is it about English that continuously reaffirms its position as a *lingua franca*, even in interdisciplinary and emerging fields like Digital Humanities? What can, on the surface, seem like a brilliant strategy for shared knowledge on a global scale actually camouflages the imposition of an intellectual blockade. In this presentation we point to strategies that could universally level the playing field for DH scholars through open access publications, quality translations, multilingual knowledge dissemination and exposure to datasets that lie outside of English. In order to illuminate these strategies in practical ways, we will provide case studies from non-English language contexts to show how we can collectively work towards greater language

inclusion and exposure in the field, an undertaking that we argue has broader consequences for social justice.

Indeed, in her work on the imperialist origins of English as a *lingua franca*, Denise Rhéaume astutely remarked that "the current usefulness of English stems from the economic, political and cultural power of, first, the British Empire, and more lately the American Empire" (2015: 151). This implies that by doing DH in English, we are (perhaps unknowingly) asserting a renewed form of scholarly imperialism that privileges Anglophone territories and long-standing, oppressive historical relationships. Breaking the cycle of Anglophone dominance means challenging the field of DH to embrace multilingualism, but, as Rhéaume argues in her piece on language politics, "obstacles to multilingualism [...] have more to do with entrenched privilege or the profit imperative than concern for the ideal conditions for democratic debate" (2015: 155). What if we, as a community of scholars, decide to change things? As we explore in this talk, challenging the *lingua franca* is a fair strategy to resist the economic, political and cultural heritage which impedes academia from moving forward in equitable ways. Scholarly just futures are possible, especially in DH, if we work to increase accessibility and promote multilingualism.

Bibliography

Cassin, Barbara. *Dictionary of Untranslatables: A Philosophical Lexicon*. Princeton, N.J: Princeton University Press, 2014.

Chan, Leslie. 'Situating Openness: Whose Open Science?' *Contextualizing Openness. Situating Open Science*, edited by Leslie Chan, University of Ottawa Press, 2019, pp. 5–22.

Desjardins, Renée. *Translation and Social Media*. London: Palgrave Pivot, 2017.

Haraway, Donna J. *Staying with the Trouble: Making Kin in the Chthulucene*. Duke University Press, 2016.

Ortega, Élika, 2014. Whispering/Translating during DH2014: Five Things We Learned. [Blog] *Élika Ortega*, Available at: <https://elikaortega.net/blog/2014/dhwhisperer/> [Accessed 8 December 2021]

Risam, Roopika. 'Colonial Violence and the Postcolonial Digital Archive'. *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy.*, Northwestern University Press, 2018, pp. 47–64, https://muse.jhu.edu/book/62714 .

Ushakov, Dmitrii. *Tolkovyi slovar' russkogo iazyka*, tom III. Moscow: Sovetskaia entsiklopediia, 1935

Zipperstein, Steven. *Pogrom: Kishinev and the Tilt of History*. New York: W. W. Norton & Company, 2018

Venuti, Lawrence. *The Translation Studies Reader*. United Kingdom, Taylor & Francis, 2000.

Cultural Landscapes in Emerging Digital Scholarship: The Search of Conceptual and Computational Frameworks

Streiter, Oliver
ostreiter@nuk.edu.tw
National University of Kaohsiung, Taiwan

Chuang, Tyng-Ruey
trc@iis.sinica.edu.tw
Academia Sinica, Taiwan

Zhan, Hanna Yaqing
hanna.yaqing.zhan@uni-hamburg.de
University of Hamburg, Germany

Hara, Shoichiro
shara@cseas.kyoto-u.ac.jp
Kyoto University, Japan

Hung, Ying-Fa
yingfa.tw@gmail.com
National Chengchi University, Taiwan

Jang, Jr-Jie
roger651017@gmail.com
JR SHEN Digital Culture Limited Company, Taiwan

Lee, Cheng-Jen
cjlee@iis.sinica.edu.tw
Academia Sinica, Taiwan

Mu, Yu-Chia Monica
monicamu@iis.sinica.edu.tw
Academia Sinica, Taiwan

Wang, Chia-Hsun Ally
allywang@iis.sinica.edu.tw
Academia Sinica, Taiwan

Wang, Yu-Huang

yuhuangwang@gmail.com
Independent Researcher, Taiwan

Panel Description

(Oliver Streiter and Tyng-Ruey Chuang)

We define cultural landscapes as **landscapes created or modified by human societies, as landscapes of historical or archaeological importance, or as landscapes chosen for an economic, spiritual, sanctuary, commemorative or other cultural function**. These landscapes are, due to their size, their internal and external heterogenity, and the process of continuous transformation, a research area that has been under-investigated in digital humanities.

In this panel we thus ask three fundamental questions. First, how can cultural landscapes be described, documented, analyzed, managed and preserved, either digitally, or through digital technologies in situ, cf. Chen and Feng (2020). Second, how can individual research or documentation efforts, conceptually or computationally, be connected to gain more holistic views of the landscape? Finally, how GIS-inspired horizontal layers can be vertically connected through linguistic or cultural descriptors?

Many aspects of cultural landscapes are complex and thus difficult to capture, e.g. in GIS-like models. These are, among others, calendric, geomantic, spiritual, and commemorative meanings of landscapes. These meanings may reside in specific geographic relations, e.g. the **fengshui** of a house, or outside the landscape, e.g. in the collective memory of a community. Where cultural practices, such as daily routines, evolve in a landscape, the calendar, the timing, and the pattern of recurrence of the practices are constitutional to practices and landscapes. Visual, olfactory, acoustic (Kopij and Pilch 2019, Manzetti 2019, Đorđević and Novković 2019), geomantic and climatic features of a landscape, in addition, require the adoption of multiple points of view for their spatial representation, e.g. wind strength as a function of time and place. A layered representation thus seems like a simplified surrogate, where e.g. a climatic function with parameters derived from multiple layers would be more adequate.

Cultural landscapes evolve in time. They can't be frozen, archived or stored in a museum and are vulnerable to disturbance and even destruction. In addition, cultural landscapes are experienced through time by a multiple of peoples in different dimensions and different research traditions with different expertise. The repeated efforts in producing documentation and data about them can span centuries, cf. Posluschny and Beusing (2019), and thus pose a real challenge in creating unified views. After all, each linking of independently produced layers relies on subjective interpretation and should not be hard-coded in the data.

A necessary but not sufficient condition for the success of the layered approach is thus the availability of shared indices and descriptors. But even if found and formulated, they can at best demonstrate spatial correlations, but not causal or cultural relations, which cannot be induced from correlations alone. Modular layers and their horizontal projections alone might thus produce only surface forms of holistic views. Yet, there are few alternatives in theory and practice as of now to represent deep structures and meanings. This panel thus proposes to bring forward ongoing works in documenting and researching cultural landscapes in East Asia by a diverse group of researchers, so as to present different approaches to cultural landscapes in digital scholarship.

The Role of Cognitive Grammars in Documenting Cultural Landscapes: Linking, mapping and interpretation

(Oliver Streiter)

In this paper we elaborate the notion of a **cognitive grammar of culture** as a systematic collection of cultural and spatial relations. Cognitive grammars might enrich or link data layers describing cultural landscapes. A grammar rule consists of a) contextual features, b) cultural entities, c) the placement of entities in context, d) a natural language description of the rule, e) motivations for the rule and f) a significance test.

In Fig. 1 a rule describes the orientation of walled gardens on the Penghu archipelago. This rule can link two layers: i) a geographic map and ii) a data set of gardens with their geo-references and orientations. By mapping areas according to rule adherence, a cultural grid is created that reflects specific gardening practices.

Alternatively, if a rule, as in Fig. 2, resembles a map, one can re-project the geographic map into the schematic map of the rule to represent the landscape from a cultural perspective. Fig. 3 shows how data layers can map geo-tagged media files to rules, triggering a cultural interpretation of the represented constellation. This can be useful in multimedia applications for educational purposes. Finally, graphical representations of rule can be manipulated in psychological experiments or interviews when investigating the notions within a community.

Cognitive grammar rules thus are a versatile representational format which through the vertical co-ocurrences of terms and indices can link various resources to analyze, interpret or visualize aspects of cultural landscapes.

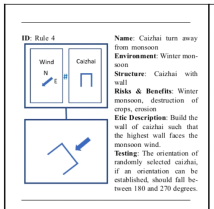


Figure 1: Walled gardens (caizhai) turn away from the wind

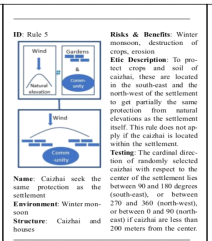


Figure 2: Caizhai, settlement and mountains

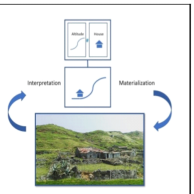


Figure 3: Mapping geo-referenced media files through elevation data and geo-references to cultural interpretations

All rules, graphs and images from Streiter, Zhan and Goudin (to appear).

Beyond GIS: Trying to make sense of Penghu cultural landscapes

(Hanna Yaqing Zhan)

Cultural landscapes are shaped by the interaction of humans with nature. This paper attempts to outline the types of cultural landscapes on the Penghu archipelago and the role that time plays for their understanding. First, we define cultural landscapes. Second, we introduce and discuss the types of cultural landscapes on Penghu. These landscapes can be religious or spiritual in nature, e.g. temples, five generals (五營) and Shigandang (石敢當) and their controlled and protected domains, or of economic interests, such as stone weirs, intertidal zones for fishing or gardens. Then we discuss the temporal aspects that are necessary for the understanding of these landscapes in their cultural context, e.g. season, monsoon, tide. Fourth, we will tackle the question, how to enrich and link documentations of cultural landscapes with these temporal data. We propose to represent a) a linear reference timeline with historic events, b) representations of generalized cyclic events, e.g. tides, monsoon, lunar circle, c) emic cyclic representations such as temple calendars, and d) generalized human activities. Activities and events are linked to landscapes and temporal representations, while all temporal representations can be linked among each other. Data on generalized human activities and emic calendars are obtained through interviews, inscriptions and, where accessible, historical resources.

Historical Landscape In the Context of Ancient Shrines

(Shoichiro HARA)

Japan is an island and mountainous country, and this topography influences the landscape and even the mentality of the Japanese people. Thus landscape is one of the critical elements to understanding the culture. As an application of GIS to humanities informatics, this paper examines the landscape of sacred sites in Japan, focusing on Shikinai-sha (式内社) in Nara Basin (奈良盆地). Nara Basin had been the seat of ancient Japanese capitols. Shikinai-sha are Japanese deity shrines recorded on *Engishiki* (延喜式). Engishiki was initially compiled from 907 to 915 to codify detailed rules for court ceremonies and protocols and is considered a precious source of historical materials. A part of Engishiki called *Engishiki-Jimmeicho* (延喜式神名帳) lists Shikinai-sha as Japanese official deities in the 10th century, which means the locations of Shikinai-sha in Nara Basin indicate sacred places in ancient times (式内社研究会, 1979). This paper first extracted geographical features around Shikinai-sha from maps (e.g. altitude and topography—flat area, foothill area, or mountain area—and numbers of mountain peaks, rivers, and slopes around a shrine), then converted these features into quantitative data, finally identify the spatial constituent elements that explain the location of shrines using some statistical analyses.

Digital Linkage of Local Knowledge: The Implementation of and Some Thoughts about the Taiwanese Religion and Folk Culture Platform (TRFC)

(Ying-Fa Hung and Jr-Jie Jang)

Taiwan is religiously one of the world's most diverse countries. This is reflected in its large number of religious buildings, networks and landscapes, as well as a large number of local records and research publications about them. However, these records and publications are notoriously difficult to use, as it is often necessary to derive basic information from the presented data. The Platform for Taiwanese Religion and Folk Culture (TRFC, website: trfc.tw) is an attempt to alleviate the problem, collating research publications and folk records on religious sites and landscapes into digital data resources. The platform is mainly based on the Taiwanese Religious Database pioneered at the Center for GIS, Academia Sinica, Taiwan, yet gradually incorporating various religious materials, routes, landscapes from related databases with the aim to establish manually or algorithmically links between them, gradually forming a local knowledge system centred on Taiwan's religious practices. An important part of the platform maps Taiwan's folk belief sites and practices, with a focus on their dynamic ritual routes, such as pilgrimages and processions. The comparative analysis, through time and regions, of data on annually recurrent events, accumulated over many years, allows researchers to understand how religious practices evolve within a religious

landscape and transform over the short or long term the cultural landscape.

Documenting Cultural Landscapes: Tools and Issues for Collaboration Across Boundaries

(Tyng-Ruey Chuang, Cheng-Jen Lee, Yu-Chia Monica Mu, Chia-Hsun Ally Wang, and Yu-Huang Wang)

Cultural landscapes are formed over time and shaped by people. The terrain of landscapes can be unwinding and embedded with artifacts of various scales. Old gravesites, abandoned factories, and seashores piled with debris, simultaneously present us with the remains of former human activity and bear witness to global changes. How to effectively document them while ensuring the documentations remain accessible to diverse communities and reusable for multiple purposes, poses many challenges.

Audiovisual and survey materials about landscapes often are coded and indexed by the time and location when they are documented. As various encodings and vocabularies are in use, reconciling documentations from multiple sources can be difficult. This problem is further compounded by the larger cultural, historical, and socioeconomic contexts inherited in the landscapes. Multiple documentations may exist yet dispersed in different collections. The annotation schemes, descriptive texts and/or semantic labels used by these collections can be heterogeneous, hence not aligned for reuse.

We view documentation materials on cultural landscapes as research datasets about which the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles apply. We propose that, with current tools and services, already landscape documentations can be made more findable, accessible and reusable, hence facilitating research collaboration. Open repositories (e.g. data.depositar.io), persistent identifiers and naming schemes (e.g. wikidata.org), data catalogue schema (e.g. DCAT), and common participatory practices (cf. citizen science) are all helpful to collection development, enrichment, and sharing.

We will show case several projects in Taiwan in documenting cultural landscapes, ranging from drone imaging for changing landscapes to preserving ephemeral street arts in a civil movement, and to folk pictures and stories about the current COVID-19 landscapes.

Bibliography

Chen, Yang and Feng, Han (2020). A digital information system for cultural landscapes: the case of Slender West Lake scenic area in Yangzhou, China. *Built Heritage*, vol. 4, article 3. https://doi.org/10.1186/s43238-020-00004-8

Đorđević, Zorana and Novković, Dragan (2019). Archaeoacoustic Research of Ljubostinja and Naupara Medieval Monastic Churches. *Open Archaeology*, vol. 5, no. 1, pp. 274-283. https://doi.org/10.1515/opar-2019-0018

Kopij, Kamil and Pilch, Adam (2019). The Acoustics of Contiones, or How Many Romans Could Have Heard Speakers. *Open Archaeology*, vol. 5, no. 1, pp. 340-349. https://doi.org/10.1515/opar-2019-0021

Manzetti, Maria Cristina (2019). The Performances at the Theatre of the Pythion in Gortyna, Crete. Virtual Acoustics Analysis as a Support for Interpretation. *Open Archaeology*, vol. 5, no. 1, pp. 434-443. https://doi.org/10.1515/opar-2019-0027

Posluschny, Axel G. and Beusing, Ruth (2019). Space as the Stage: Understanding the Sacred Landscape Around the Early Celtic Hillfort of the Glauberg. *Open Archaeology,* vol. 5, no. 1, pp. 365-382. https://doi.org/10.1515/opar-2019-0023

式内社研究会 (Research Group of Shrines Listed in the Engishiki) (1979). *The Report of the Shikinaisha Survey*, Kogakkann University Press (Japanese).

Long Presentations

A Computational Approach to Epistemology in Poetry of the Long Eighteenth Century

A Case Study in Objects and Ideas

Algee-Hewitt, Mark Andrew
malgeehe@stanford.edu
Stanford University, United States of America

Poetry occupies a unique position in the computational study of literary text. Most recent computational linguistics work on poetry has focused on the analysis of meter or text generation (e.g. Lau et. Al 2018; Hämäläinen 2018). In cultural analytics, studies of poetry have focused primarily on sound studies (MacArthur et. Al. 2018) or meta-analyses of the context of its performance (Basnet and Lee 2021). In this paper, I leverage the unique textual and linguistic properties of poetry to explore how ideas are communicated poetically and whether this communication changes over time within a historical corpus in ways that respond to the shifting epistemologies of the Enlightenment and post-Enlightenment approaches to knowledge production. Given the interest in empiricism during the early eighteenth century, which shifts to an idealism-based philosophy during the early nineteenth century, I investigate whether the use of object-nouns (nouns representing material artifacts) and concept-nouns (nouns representing concepts) in poetry from this period was influenced by this change in how knowledge is produced and communicated.

Corpus

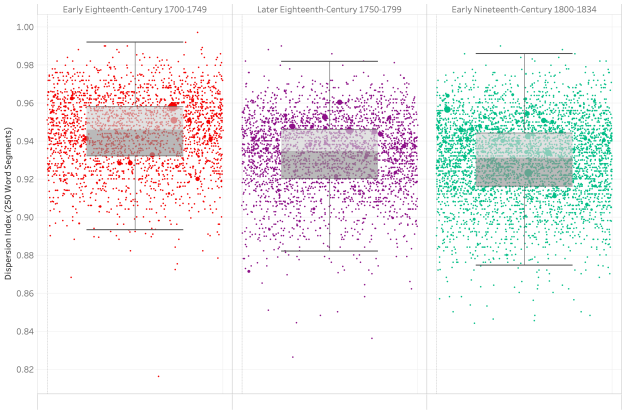
For this project, I use the Chadwyck-Healey poetry archive, published by ProQuest. The final filtered corpus includes 55,293 poems written in English between 1700 and 1840.

Feature Selection

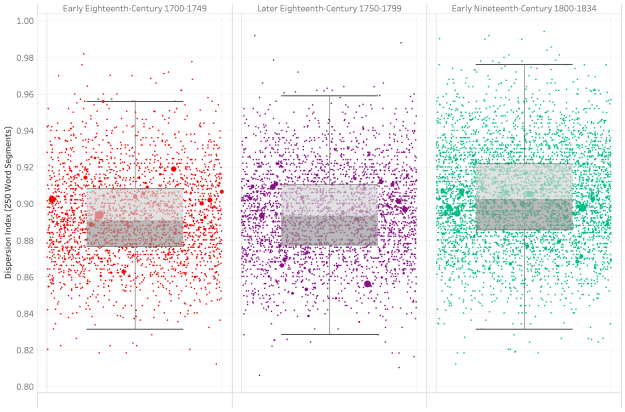
Previous work on concrete and abstract features has used word embeddings to find related terms (Heuser 2017), hand-curated lists of associated terms (Heuser and Le-Khac 2012), or found words with Latinate and Anglophone roots (Underwood 2012). For this project I also adopt embeddings to identify object-nouns and concept-nouns, but use a supervised learning approach. Beginning with a pre-trained GloVe word embedding model on the common core, consisting of 840 billion tokens representing a 1.2 million word vocabulary embedded within 300 dimensions (Pennington et. Al. 2014), I use the embedding dimensions as features to create a machine learning model to classify nouns into object and concept categories. Working from a full NLP dependency parse of the poetry corpus (coreNLP) to filter the embedding model for words used as nouns in eighteenth and nineteenth-century poetry, I trained a linear discriminant analysis to identify object and concepts words based on a 50 word seed list, retaining only those terms classified with a 75% or higher probability of being either an object (4418 nouns) or a concept (4843 nouns).

Analysis

To explore the relationship between the use of object-nouns and concept-nouns in poetry across the long eighteenth and early nineteenth centuries, I first assess whether there is a difference between how these two classes of words are distributed within the poems of the period and whether this relationship changes over time. To compare the distribution of each word set, I calculated the dispersion index (variance/mean) for each word set in each poem (figure 1).



a. object words

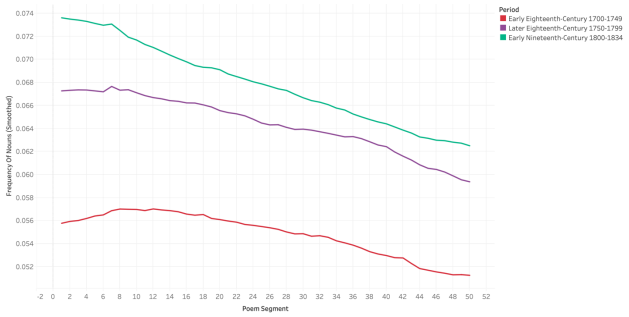


b. concept words

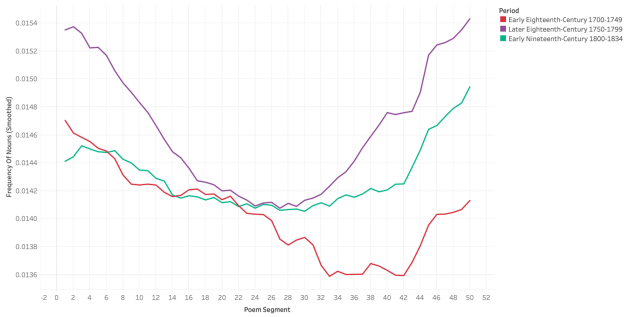
Figure 1:
a. dispersion index of object words in poems written between 1700 and 1834; b. dispersion index of concept words in poems written between 1700 and 1834 (individual poems are jittered on the x-axis for legibility within each period)

As the figures indicate, object words evidence significantly (pairwise Wilcoxon p-value of 2.2 e-16; Cohen's d of 0.8) lower dispersion in poems written earlier in the eighteenth century than in the later eighteenth or nineteenth centuries. Conversely, concept words are much more significantly distributed (p-value 2.2 e-16, Cohen's d of 0.7) throughout earlier poems. In other words, poems written during the Enlightenment distribute concept words throughout the poem, but cluster objects together, while poems written in the early nineteenth century cluster object words less, but cluster concept words to a greater extent. This potentially offers evidence that eighteenth-century poetic epistemology relied on empirical evidence (representing clusters of direct objects) more than nineteenth-century poetry.

To explore how these clusters playout within the poems themselves, I aggregated the number of object and concept nouns within each 50 th (using overlapping windows) of poems written during each of the three periods (figure 2).



a. object nouns



b. concept nouns

Figure 2:
Aggregate position of a. object nouns and b. concept nouns within poems written during the early eighteenth century, the later eighteenth century and the early nineteenth century

Once again, the differences between both object and concept nouns, as well as between each time period, is apparent. Across all three time periods, object nouns fall across the poems, while in the two later periods, concept nouns are strongly clustered at the beginning and the end of their respective poems. This effect is similar for early eighteenth-century poetry, however the fall is much more dramatic, and the nadir occurs later in the poem. This suggests that concept words cluster within the frame of the poems from later periods, while they are more evenly used throughout poems from the earlier period.

As a final step, I calculate the most frequent verbs associated with all objects and all ideas (extracted from a dependency parse of the poetry). These associated verbs, I argue, can aid us in understanding how the changing dispersal and frequency of both ideas and concepts correlate with their use in the poems. The data reveals that across the eighteenth century objects become less the objects of "doing" and "knowing," and more the objects of "seeing," "hearing," and "feeling." Concept nouns remain more stable in relation to "thinking" and "knowing," but become the objects of "being."

Conclusions

The quantitative changes of the dispersal of nouns representing objects and ideas, as well as their average frequency across poems, refute the standard narrative of poetic evolution that emphasizes the object-oriented nature of Romantic period poetry. Instead, following the science-oriented epistemology of the Enlightenment, eighteenth-century poems used organized clusters of objects in order to inductively develop and support emergent ideas, while in Romantic period poetry, objects are scattered throughout the poems in support of the centralized and organized ideas that animate the poem.

Bibliography

Basnet, Anik and James Jaehoon Lee. "A Network Analysis of Postwar American Poetry in the Age of Digital Humanities." *Journal of Cultural Analytics* 4(2021): 180-217.

Hämäläinen, Mika. "Poem machine-a co-creative nlg web application for poem writing." *The 11th International Conference on Natural Language Generation Proceedings*

of the Conference. The Association for Computational Linguistics, 2018.

Heuser, Ryan "Word Vectors in the Eighteenth Century" ADHO Conference, 2017.

Heuser, Ryan and Long Le-Khac. "A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method." *Pamphlets of the Stanford Literary Lab* 4(2012).

Lau, Jey Han, Trevor Cohn, Timothy Baldwin, Julian Brooke, and Adam Hammond. "Deep-speare: A joint neural model of poetic language, meter and rhyme." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics 2018.

MacArthur, Merit, Georgia Zellou, and Lee M. Miller. "Beyond Poet Voice: Sampling the (Non) Performance Styles of 100 American Poets." *Journal of Cultural Analytics* (2018): np.

Pennington, Jeffery, Richard Socher, and Christopher D. Manning. *GloVe: Global Vectors for Word Representation.* 2014.

Underwood, Ted. "Etymology and nineteenth-century poetic diction; or, singing the shadow of the bitter old sea." *The Stone and the Shell* 2012.

Gender Assignment as an Event: a contemporary approach to adequately depict historical gender categories

Andrews, Tara Lee
tara.andrews@univie.ac.at
Institute for History, University of Vienna; Austrian Centre for Digital Humanities and Cultural Heritage, Austrian Academy of Sciences

Ebel, Carla
carla.ebel@oeaw.ac.at
Austrian Centre for Digital Humanities and Cultural Heritage, Austrian Academy of Sciences

Deierl, Marin
marius.deierl@univie.ac.at
Institute for History, University of Vienna

When we consider the historical and contemporary reality of gender identities, we frequently find cases where gender assignments change over a person's lifetime. According to the discourse on gender in the last 70 years (Beauvoir, 1949; Butler, 1990; Lugones, 2007) it is recognised that gender is not an innate attribute of a person, but a categorization attributed according to regional and historical circumstances, usually based on physical features. Attribution processes and the available gender vocabulary can vary over time and place, although gender categorization is usually performed in temporal proximity to a person's birth.

The nature of this categorization, however, is not adequately represented in existing data standards. Although modeling gender is not addressed in the Cidoc CRM 7.1.1 documentation (Bekiari et al., 2021) the use of "P2 has type" and "E55 Type" is the default solution, which treats gender as an untemporalized type that is assigned to a person. [1] Other models for biographical data, such as schema.org [2] and the data model of the German National Library [3] also treat gender as an inherent and unchanging category. Although the FOAF Vocabulary Specification claims not to (Brickley and Miller, 2000), the FOAF model does not include a general means of recording change in personal information over time. In this paper we present an approach, based on the CIDOC-CRM and its principles of event modelling, that moves beyond static gender roles to encompass assignment of gender identity as an event in people's lives. [4]

The test bed for our approach is the RELEVEN project. RELEVEN focuses on the so-called "short 11th century" in the eastern Christian world, spanning the territory from Italy and the Balkans through to Iran; the project explores the connections between people, regions and ideas across the cultures that made up the Christian world, especially in its eastern areas where the majority of people lived and in the areas of heaviest interregional interchange with the Muslim world. In this context we find three genders that played a significant role in the historical sources for this period: "eunuch", "female" and "male"; in line with the wider scholarship on gender, it has long been acknowledged that, not only within the Byzantine world but also in the Islamic world and elsewhere in Asia, they were indeed considered a third gender (Ringrose, 2003) and they are represented as such in existing databases such as the Prosopography of the Byzantine World (Jeffreys et al., 2017).

While the first naive approach might be to classify eunuchs as males who underwent a castration event, this would not reflect the historical situation appropriately – in the sources there are some eunuchs specified as "congenital eunuchs". While it isn't entirely clear what constituted a "congenital eunuch", the suggestion seems to be that in these cases no castration event ever occurred; whether the persons concerned were intersex or did not form gender-typical physical features by the time of puberty, it seems that they were recognised relatively early in life as belonging to the "eunuch" category rather than the "male" category. Thus, for any particular eunuch, unless it is specified when and how they came to be assigned to the eunuch state, we

cannot know whether it was a "congenital" or inherent classification, or an explicit change of status that was forced with mutilation.

Given this gap in the evidence that we have, it becomes clear that we must also treat 'female' and 'male' as gender assignments rather than sex assignments; after all, almost all of the evidence we have concerns the social roles of the people involved rather than their anatomy or genetics. While these would have aligned in the expected way in the vast majority of cases, it makes more sense to use the classification for which we actually have written evidence: namely, the social construct. From this perspective, it follows that we cannot map the eunuch category as a special case of a binary; as argued by Ringrose (2003), if we are modeling the social construct (gender) rather than the biological characteristics (sex), we must represent the three categories we encounter on an equal footing: an assignment that usually does not, but nevertheless can, change over time.

In the RELEVEN data model, all gender assignments are treated as events, and specifically as E17 Type Assignment events in the CIDOC-CRM ontology. [5] The usual assumption is that, unless specified otherwise in a source, these assignments happen at birth or at the latest by the age of puberty; since the assignment is a subclass of E7 Event, the association of a date or date range to the event becomes straightforward. The relevant classes and properties from Cidoc CRM 7.1.1. are given in OWL in Figure 1.

```
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix crm: <http://www.cidoc-crm.org/cidoc-crm/> .

crm:E17_Type_Assignment rdf:type owl:Class .

crm:E1_CRM_Entity rdf:type owl:Class .

crm:E21_Person rdf:type owl:Class .

crm:E39_Actor rdf:type owl:Class .

crm:E52_Time-Span rdf:type owl:Class .

crm:E55_Type rdf:type owl:Class .

crm:P14_carried_out_by rdf:type owl:ObjectProperty ;
                                                rdfs:domain crm:E17_Type_Assignment ;
                                                rdfs:range crm:E39_Actor .

crm:P17_was_motivated_by rdf:type owl:ObjectProperty ;
                                                   rdfs:domain crm:E17_Type_Assignment ;
                                                   rdfs:range crm:E1_CRM_Entity .

crm:P41_classified rdf:type owl:ObjectProperty ;
                                              rdfs:domain crm:E17_Type_Assignment ;
                                              rdfs:range crm:E21_Person .

crm:P42_assigned rdf:type owl:ObjectProperty ;
                                            rdfs:domain crm:E17_Type_Assignment ;
                                            rdfs:range crm:E55_Type .

crm:P4_has_time-span rdf:type owl:ObjectProperty ;
                                               rdfs:domain crm:E17_Type_Assignment ;
                                               rdfs:range crm:E52_Time-Span .
```

Fig. 1:
An OWL representation of the relevant CIDOC-CRM classes and properties for a Gender Assignment Event

In Figure 2 is a schematic example of an assertion in our data model, claiming on the basis of the Chronicle of Michael the Syrian that the son of the last Doukas emperor, Michael VII, was castrated by the usurper Nikephoros Botaneiates during his reign. This approach to modeling, taken together with the wider context of RELEVEN's assertion-based (rather than fact-based) data model (Baillie et al. 2021), means that we can also represent conflict or confusion in the sources about a person's gender identity, where this conflict arises, simply by adding a competing assertion based on another source.

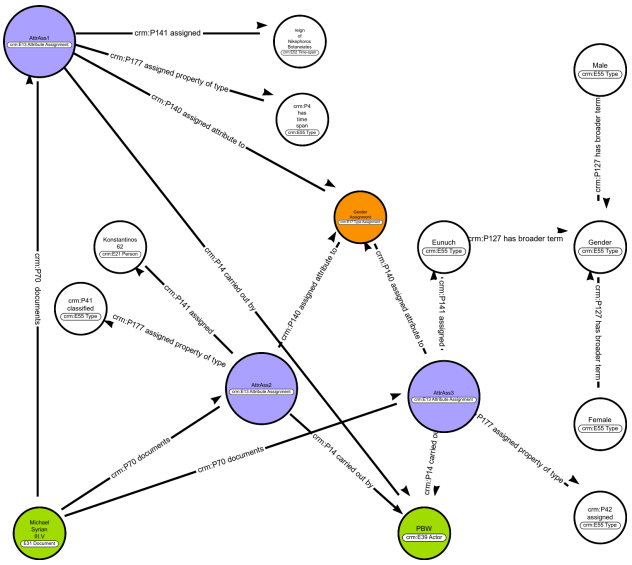


Fig. 2:
Assertion that Constantine Doukas became a eunuch between 1078–81

We present our approach in the hope that it will not only be useful in the context of our own project, but that its applicability to other prosopographical data sets in historical and geographical contexts can be recognised. The Hijras of South Asia and the *burrneshat* of Albania indeed often acquire their gender status via an event (a ritual or the taking of an oath), making this a very suitable model for these cases; transgender people all over the world can also easily be represented this way. The model is also relevant to post-mortem gender attributions, such as those often made in the context of archaeological excavations.

Acknowledgments

This work arises from a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 101002357).

Bibliography

Baillie, J., Andrews, T. L., Romanov, M., Knox, D. and Vargha, M. (2021). Modelling Historical Information with Structured Assertion Records. *Digital History Berlin*. Berlin https://dhistory.hypotheses.org/518 (accessed 10 December 2021).

Beauvoir, S. de (1949). *Le deuxième sexe*. Gallimard.

Bekiari, C., Bruseker, G., Doerr, M., Ore, C.-E., Stead, S. and Velios, A. (eds). (2021). Definition of the CIDOC Conceptual Reference Model https://www.cidoc-crm.org/Version/version-7.2.

Brickley, D. and Miller, L. (2000). FOAF Vocabulary Specification http://xmlns.com/foaf/spec/ (accessed 10 December 2021).

Butler, J. (1990). *Gender Trouble: Feminism and the Subversion of Identity*. Routledge.

Jeffreys, M. (2017). *Prosopography of the Byzantine World, 2016*. King's College London https://pbw2016.kdl.kcl.ac.uk/.

Lugones, M. (2007). Heterosexualism and the Colonial / Modern Gender System. *Hypatia*, **22**(1). Indiana University Press: 186–209.

Ringrose, K. M. (2003). *The Perfect Servant : Eunuchs and the Social Construction of Gender in Byzantium*. (ACLS Humanities E-Book Series). Chicago: University of Chicago Press https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=212703&site=ehost-live (accessed 10 December 2021).

Notes

1. The CIDOC-CRM model originally included an explicit way to model gender, but the relevant definitions were removed in 2001 on the grounds that "there is nothing more important about gender than about any other properties giving rise to a set of people". See comments on the closed issue: https://www.cidoc-crm.org/Issue/ID-38-delete-gender and recommendations in the SARI documentation for example: https://docs.swissartresearch.net/et/persons/.
2. https://schema.org/gender
3. https://d-nb.info/standards/elementset/gnd#gender
4. We are aware that our current model can not represent gender in all its complexicity. A notable example that captures more and other dimensions of gender is the "GenderedCHContents" ontology (Bikakis and Kyvernitou 2017). Based on our data and RELEVENs focus on assertions, we see the model respresented in the upcoming paper as an adequate approach for our project.
5. This is an uncontroversial modeling decision, since 'P2 has type' is a shortcut of the event-based path E1 CRM Entity : P41i was classified by : E17 Type Assignment; E17 : P42 assigned to : E55 Type. See https://www.cidoc-crm.org/Property/P2-has-type/version-7.1.1.

Online Readership and Perceptions of Genres Over Time

Antoniak, Maria
maa343@cornell.edu
Cornell University

Walsh, Melanie
melwalsh@uw.edu
University of Washington

Mimno, David
mimno@cornell.edu
Cornell University

When today's readers think about "science fiction," what kinds of books do they think of? Do they think of Ursula K. Le Guin's 1969 novel *The Left Hand of Darkness*, or Neal Stephenson's 1992 *Snow Crash*, or Hugh Howey's 2011 *Silo* series? Another way of phrasing this question might be: When today's readers think about "science fiction," what *era* of books do they think of? Do they think of the 1960s, the 1990s, the 2010s, or all of the above? What about "romance," "fantasy," "classics," and "vampires"? What historical eras do readers imagine when they think about these categories, and how do they compare to science fiction?

Data from online reading communities like Goodreads and LibraryThing — where readers rate, review, and categorize books — enables us to answer these questions and to explore the relationship between genre and historical period in the minds of readers. We specifically examine reception data from LibraryThing, where users can add any number of free-text tags to any book and where all of this data is publicly available (unlike Goodreads, where review data is mostly hidden (Walsh and Antoniak 2021)) and we combine it with book publication information in order to better understand which eras of books readers are categorizing with which tags. For example, it turns out that the median publication date for books tagged as "science fiction" is the year 1989, which is significantly earlier than the median publication date for the tag "vampires," the

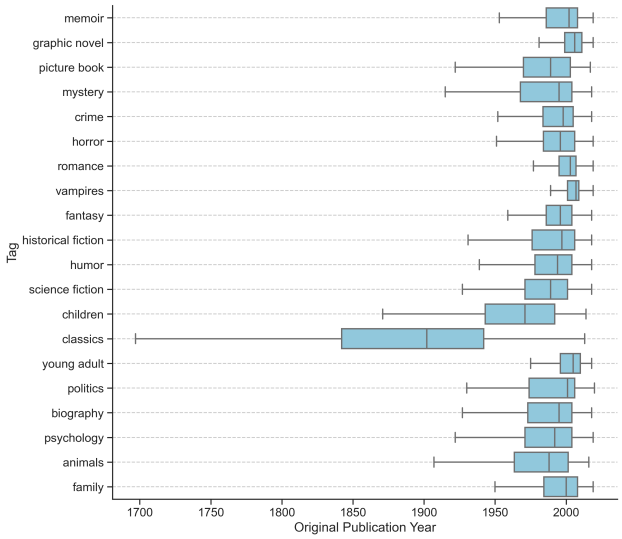
year 2007. This information can help us identify emergent genres as well as understand how readers perceive genres historically.

This work builds on recent studies of online reading communities that take advantage of the scale and diversity of online data to examine readers' preferences (Manshel et al., 2019; Bourier and Thelwall, 2020; English et al., 2021). In particular, our work adds to explorations of online readers' perceptions of genres (Hegel, 2018; Antoniak et al., 2021; Walsh and Antoniak, 2021), which have used natural language processing methods on reviews to measure affinities and differences between user-applied tags. By focusing on publication dates, we contribute a new layer in understanding how today's readers perceive literary genres.

Comparing the Top 20 Most Popular Tags By Publication Date

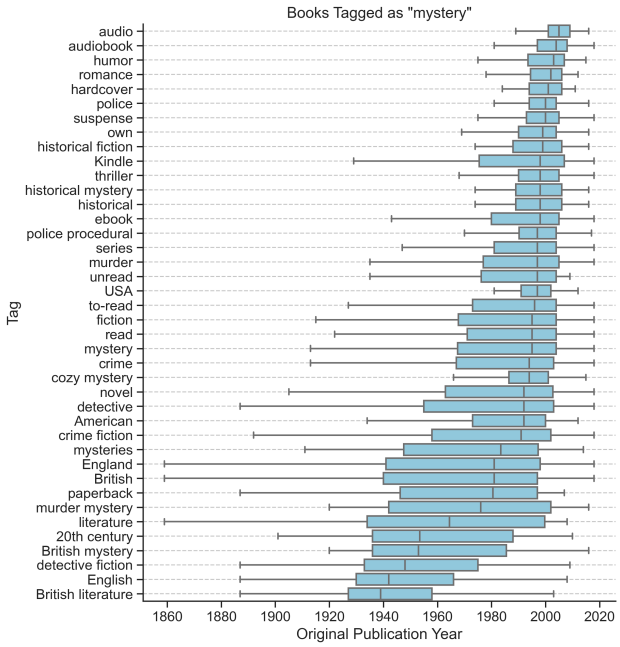
We begin by examining the distribution of publication years for the 20 most popular tags. From the set of most popularly used tags on LibraryThing, we select the 20 most popular tags that are not meta (e.g., *to-read*) or repetitive (e.g., *classic* when already including *classics*), which ensures that the selected tags more closely resemble genre or subject labels. For each of these 20 tags, we scrape the metadata of the 1,000 books most often assigned that tag. This metadata includes the book's title, author, original publication year, user ratings, user reviews, and the full set of tags that users have applied to this book. Of the tags assigned to any book, we include only those assigned by at least 10 users.

The resulting distributions suggest that users perceive certain tags as belonging to earlier historical periods than others. For example, the median publication date for *classics* (1900), *children* (1971), and *picture book* (1989) are all decades earlier than *graphic novel* (2006), *vampires* (2007), or *young adult* (2005). This contrast likely points to the fact that *graphic novel*, *vampires*, and *young adult* are more recent, emergent genres. The publication distribution for *horror*, a genre that sometimes includes characters who are vampires, is also much earlier and wider than *vampires*, again suggesting that *vampires* is its own distinct, historically-specific genre. Unlike vampires, the tags *mystery* and *science fiction* both have wide publication distributions that begin in the early 20th century.

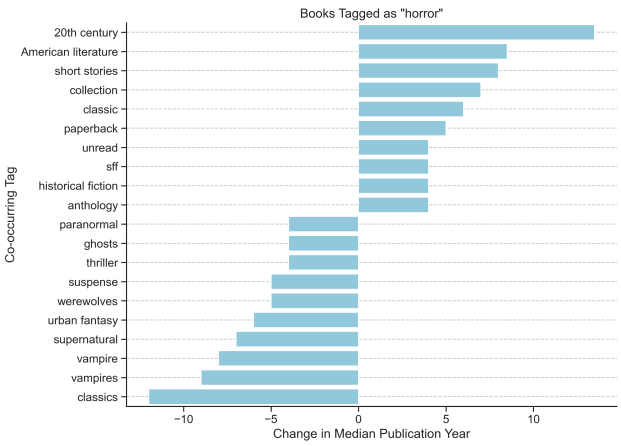


Exploring Each Tag By Co-Occurring Tags and Publication Date

We then consider each of the most popular 20 tags in turn, examining each tag's 40 most commonly *co-occurring tags* and their corresponding publication date distributions. This view allows us to see at an even more granular level which aspects of a tag might contribute to its perceived relationship with different historical eras. For example, *mystery* books that receive the tag *USA* are, on average, newer than *mystery* books that receive the tags *England* or *British*. Similarly, *mystery* books that receive the *hardcover* tag are typically newer than *mystery* books that receive the *paperback* tag, likely because new books are published in hardcover before they are published in paperback. These co-occurring tags and publication date distributions give us a more multidimensional understanding of how readers use the tag *mystery*.



We also examine the difference between the median publication year for a tag over all the books and the median publication year for only those books most often assigned to one of our 20 target tags. For example, for the books most often receiving the tag *horror*, the co-occurring tag *20th century* is assigned to books with more recent publication years compared to the full set of books receiving the *20th century* tag. But when *horror* books receive the *vampires* or *werewolves* tags, these tend to have older median publication dates than they normally would. The *horror* genre, according to LibraryThing users, consists of newer books tagged *20th century*, *American literature*, and *short stories* as well as older books tagged *classics*, *vampires*, and *supernatural* books, in comparison to the other genres.



Conclusion

We have shown that tags and book publication dates can be combined to give us multidimensional views of how readers use tags in online reading communities. Of course readers' use of tags and their perceptions of genre are shaped by a multitude of economic and sociological factors (McGrath, 2020) that we do not address here. And we believe that considering these factors and attendant critical scholarship would help draw out the significance of our findings even more. While we leave this specific synthesis to future work, we conclude with the claim that reception data like tags and data-driven research like our study can contribute to ongoing conversations about literary genre and help illuminate how contemporary readers think about and interact with books.

Bibliography

Antoniak, M., Walsh, M., and D. Mimno. (2021). "Tags, Borders, and Catalogs: Social Re-Working of Genre on LibraryThing." Proceedings of the ACM on Human-Computer Interaction 5. CSCW (2021): 1-29.

Bourrier, K., and Thelwall, M. (2020). "The social lives of books: Reading Victorian literature on Goodreads." Journal of Cultural Analytics 1.1: 12049.

English, J. F., Enderle, S., and Dhakecha, R. (2018). "Mining Goodreads: Literary Reception Studies at Scale," https://pricelab.sas.upenn.edu/projects/goodreads-project, accessed December 10, 2021.

Hegel, A. (2018). Social Reading in the Digital Age. University of California, Los Angeles.

McGrath, L. (2020). "America's Next Top Novel." Post45.

Walsh, M., and Antoniak, M. (2021). "The Goodreads "Classics": A Computational Study of Readers, Amazon, and Crowdsourced Amateur Criticism." Journal of Cultural Analytics 4: 243-260.

Manshel, A., McGrath, L.B., and Porter, J. D. (2019). "Who Cares about Literary Prizes?." Public Books 3.

Manifesting the manifesto: DH and the climate crisis

Baillot, Anne
anne.baillot@ens-lyon.fr
Le Mans Université

Gil Fuentes, Alexander
agil@columbia.edu
Columbia University

Glover, Kaiama L
kglover@barnard.edu
Barnard College

Peaker, Alicia
apeaker@barnard.edu
Barnard College

Roeder, Torsten
dh@torstenroeder.de
Bergische Universität Wuppertal

Scholger, Walter
walter.scholger@uni-graz.at
Universität Graz

Walton, Jo Lindsay
j.c.walton@sussex.ac.uk
University of Sussex

In 2021, an international group of DH scholars drafted the manifesto "DH and the Climate Crisis". It argues that the climate crisis is bound to move the lines not only of DH research and teaching practices, but also of the self-understanding of the discipline.

The digital is material. As digital humanists, every project we create, every software application we use, every piece of hardware we purchase impacts our environment. In her 2014 DH keynote address, Bethany Nowviskie exhorted us to "attend to the environmental and human costs of DH." As a field, we work adjacent to fields whose research relates to environmental crises: to science and technology studies scholars concerned with their relations to Land/land (Liboiron, 2021); to archivists describing how their approaches to digital preservation are environmentally unsustainable (Pendergrass et al., 2019); to artificial intelligence ethics scholars investigating the intersectional harms of large language models, some of whom have been fired by big tech for speaking out, others of whom only feel able to speak anonymously (Bender et al., 2021); to historians quantitatively analysing the disinformation tactics of big oil (Supran and Oreskes, 2021). Yet climate justice and environmental impacts remain under-researched in the field of digital humanities and underrepresented in our conferences and literature.

In response to this need for greater awareness, an international group of digital humanities scholars and practitioners gathered virtually to share knowledge and experiences and build momentum. These virtual meetings resulted in the drafting of the manifesto "DH and the Climate Crisis," which was published in the summer of 2021. The group's shared values around open access, open collaboration, minimal computing, and environmental responsibility drove decisions about what form the draft would take, where and how it would be published, and how our community could shape its directions.

The publication of a first, already collaborative version of the manifesto text was met with interest and annotated by community members over the course of several weeks. Some commenters expressed scepticism about the role of DH as a research field in relation to the climate crisis, others about the form of the manifesto. These comments sparked vibrant discussions and led to the publication of the current version. The goal of our paper is to foster stronger awareness and concrete initiatives regarding the consequences of the environmental impacts of research and teaching practices throughout the DH community at large. We argue that the climate crisis is bound to move the lines not only of DH research and teaching practices, but also, at an epistemological level, of the self-understanding the discipline has of itself.

In the first part of the paper, we will present the major arguments of and for the manifesto. It can hardly be denied that Digital Humanities research is in many regards more resource-intensive than most of the Humanities disciplines. But this is not the only reason why the community should feel particularly concerned: DH is also positioned to recognize, describe, and tackle the practices across the Humanities that are carbon intensive and resource extractive, and in that sense it has the potential to engage a new approach to research and teaching in an age of climate crises. We will also argue that Digital Humanities have elaborated their self-understanding based on a deconstructive approach of the canon in which the computational analysis of large data sets is a significant factor (Moretti 2000), and that rethinking research practices according to a minimization of environmental impact - in a context where growth remains a core value - is an opportunity to reframe our understanding of the originality of our approach.

In the second part of the paper, we present a series of actions and initiatives that have emerged following the publication of the "DH and the Climate Crisis" Manifesto. The first consists of a "Next Steps" document, linked to the manifesto, where we are collectively sharing the disparate, but interlinked, actions that we are taking in our various institutions, sharing experiences and best practices. Second, we will present environmental initiatives emerging in Europe that aim at fostering communities of action,

and supporting one another to take practical actions both immediately and in the longer term. Throughout this second part of the paper, we will highlight connections between the local and the global, and explore how DH communities around the world can collaborate to address planetary challenges.

The next few years are crucial if the world is to meet the target of limiting global warming to 1.5 degrees. To work effectively on such a timescale, our practice will need to evolve rapidly. We therefore expect that our paper in July may also reflect new challenges, opportunities, collaborations, and points of interest to emerge in the first half of 2022, which are not yet known to us. Through this manifesto and paper, we aim to find and build resilient communities within digital humanities—to connect, encourage, and support the ways we are already responding to the global climate crisis, and all the ways to come.

Bibliography

Bender, Emily M, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?', 2021, 14. https://doi.org/10.1145/3442188.3445922.

Liboiron, Max. Pollution Is Colonialism. Durham, N.C.: Duke University Press, 2021.

Moretti, Franco. Conjectures on World Literature. New Left Review (I. Jan/Feb 2000): https://newleftreview.org/issues/ii1/articles/franco-moretti-conjectures-on-world-literature

Nowviskie, Bethany. 'Digital Humanities in the Anthropocene'. Digital Scholarship in the Humanities 30, no. suppl_1 (1 December 2015): i4–15. https://doi.org/10.1093/llc/fqv015.

Pendergrass, Keith, Walker Sampson, Tim Walsh, and Laura Alagna. 'Toward Environmentally Sustainable Digital Preservation'. The American Archivist, June 2019. https://doi.org/10.17723/0360-9081-82.1.165.

Supran, Geoffrey, and Naomi Oreskes. 'Rhetoric and Frame Analysis of ExxonMobil's Climate Change Communications'. One Earth 4, no. 5 (21 May 2021): 696–719. https://doi.org/10.1016/j.oneear.2021.04.014.

DFG 3D-Viewer – Development of an infrastructure for digital 3D reconstructions

Bajena, Igor Piotr
igorpiotr.bajena@unibo.it
Hochschule Mainz – University of Applied Sciences, Germany; University of Bologna, Italy

Dworak, Daniel
daniel.dworak@hs-mainz.de
Hochschule Mainz – University of Applied Sciences, Germany

Kuroczyński, Piotr
piotr.kuroczynski@hs-mainz.de
Hochschule Mainz – University of Applied Sciences, Germany

Smolarski, René
rene.smolarski@uni-jena.de
Friedrich-Schiller-Universität Jena, Germany

Münster, Sander
sander.muenster@uni-jena.de
Friedrich-Schiller-Universität Jena, Germany

Introduction

An important element in digital 3D reconstruction, in the fields of archeology, art and architecture history, is the subsequent visualization of the result (Messemer, 2016). The standardization of the documentation and publication is seen as the most important priority across the board (Cieslik, 2020). Widely established 3D repositories with integrated 3D visualization such as Sketchfab (https://sketchfab.com/) belong to a commercial offer, while 3D viewers introduced by scientific institutions like Kompakkt (https://kompakkt.de/home) or by other research projects like patrimonium.net (Dworak, Kuroczyński, 2016) have still not provided approved and applied standards for the documentation and publication of 3D models in the field of hypothetical 3D reconstruction of art and architecture.

Project assumptions

Against this background, the project "DFG 3D-Viewer - Infrastructure for digital 3D reconstructions" was launched, which the goal is to provide an offer of permanent infrastructure for decentralized web-based display of models in the DFG 3D-Viewer and in suitable Virtual Research Environments (VRE), accompanied by low threshold interface usage (http://dfg-viewer.de/en/dfg-3d-viewer). The work presented here concerns the results of the first phase

of the project including the definition of documentation standards for web-based 3D publication of the digital reconstruction models and the development of the web-based 3D-viewer for digital datasets as a minimal-effort plugin. Proposed solution considers a generic approach with adaptability and reusability (Münster, 2019), respects FAIR principles (https://www.go-fair.org/fair-principles/) and follows existing DFG (German Research Foundation) standards (https://www.dfg.de/en/research_funding/principles_dfg_funding/good_scientific_practice/index.html).

Minimal documentation standard

Analysis of documented metadata of the chosen commercial and institutional 3D repositories formed the basis for the definition of a scheme for documentation (Fig. 1). The developed data set was discussed among the community in the form of a survey, which significantly advanced the work towards establishing a standard. It also allowed to emerge documentation-related functionalities of the viewer, such as automatic rendering of the preview images or the displaying the information about model geometry (3D metadata) in the viewer window. The documentation scheme was implemented in a new prototypical 3D repository created in WissKI-based VRE (http://wiss-ki.eu/), which has already been successfully used in several projects of digital reconstructions at the University of Applied Sciences Mainz (Kuroczyński et al., 2022; https://www.new-synagogue-breslau-3d.hs-mainz.de). The data model in the repository uses the CIDOC Conceptual Reference Model (https://www.cidoc-crm.org/) as an ontology. The fundamental research on data modelling was carried out along the community in order to concerns about different combinations of classes and properties to describe the same aspects of documentation.

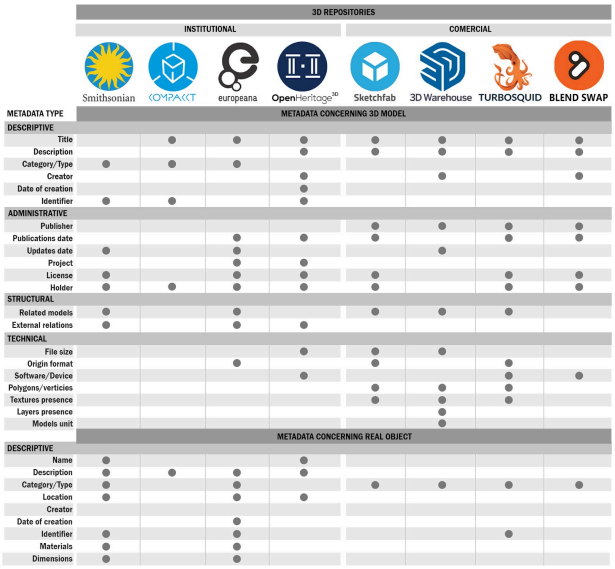


Fig.1
The comparison of metadata sets in chosen institutional and commercial 3D web-based repositories (©2021, Hochschule Mainz).

Framework architecture of the 3D Viewer

Comparing present 3D viewer solutions, it was decided to take the following properties into account: support for 2D & 3D objects, variety of source formats, support for complex objects, modern technology based, suitable for hand-modeled and laser-scanned objects, 3D world operations, level of detail (LoD) as models representations, compression of 3D objects, 3D metadata, utilities/tools, documentation.

It appears that only a few 3D viewers fulfill more than half of the requirements. In fact, some of the analyzed applications support 2D/3D objects and a variety of formats, but some are still missing (PLY, XYZ, DAE) (Champion, Rahaman, 2020). These technologies are optimized for hand-modeled objects, while others only for laser-scanned ones. Three of them allow 3D world operations and support 3D metadata, nevertheless none of them supports 3D compression.

	3DHOP	Smithsonian	Kompakkt	Sketchfab	Inception	Media Wiki	DFG-3D-Viewer
Support of 2D & 3D objects	o	+	+	o	o	o	+
Variety of formats	-	o	+	+	-	-	+
Complex objects	o	+	o	o	+	N/A	+
Modern technology based	-	+	+	+	+	N/A	+
Hand-modelled objects	-	-	+	+	-	+	+
Laser-scanned objects	+	+	o	-	+	N/A	+
3D world operations	+	+	+	+	-	-	+
LoD (models representations)	+	-	-	o	-	-	o
Compression of 3D objects	+	-	-	-	-	-	+
3D metadata	+	+	+	+	-	-	+
Utilities, tools	+	+	-	+	-	-	+
Documentation	+	+	-	+	-	-	o

+ property is supported; - property is not supported; *o* property is partly supported; *N/A* information not available

Fig.2
Comparison of functionalities of the most competitive 3D viewers on the market (©2021, Hochschule Mainz).

The architecture of the DFG 3D-Viewer is developed considering existing web-based 3D viewers (Champion, Rahaman, 2020; Fernie at al., 2020). Conducted research compares existing infrastructures and viewers (Fig. 2), as well as the concept of a modular architecture for the DFG 3D-Viewer. It concludes that the framework for the scientific 3D infrastructure (considering documentation and publication) should be cross-browser, platform independent and based on modern, promising and long-term supported technology. The viewer should allow viewing of 3D models with textures, stored in the most common formats used nowadays , i.e. OBJ, DAE, FBX, JSON (Cieslik, 2020). It should be also capable of loading 2D images (JPG, PNG, TIFF) (Cieslik, 2020), 3D metadata and provide 3D world operations on models (Fernie at al., 2020). Solution should be integrable out of the box, open source and client-only in order to distribute workload away from the server and minimize the requirements for repository providers to support the DFG 3D-Viewer.

The developed framework is based on the existing 3D library - three.js. Implementation was prepared in modern and interchangeable programming languages and technologies as JavaScript, PHP or Python. Architecture is optimized to be technology-independent and can be easily exchanged for any other client-side viewer. The solution is suitable for complex, hand-modeled, laser-scanned objects and 3D metadata as well. The viewer is extended to meet the requirements of the specialist community, including the possibility of displaying highly complex geometries and multiple data formats (inter alia IFC and FBX) (Fernie at al., 2020). Moreover, uploading 3D data triggers automatic unattended compression, based on Draco algorithm, and encoding into the glTF format which is optimized for web-based visualization (Fig. 3).

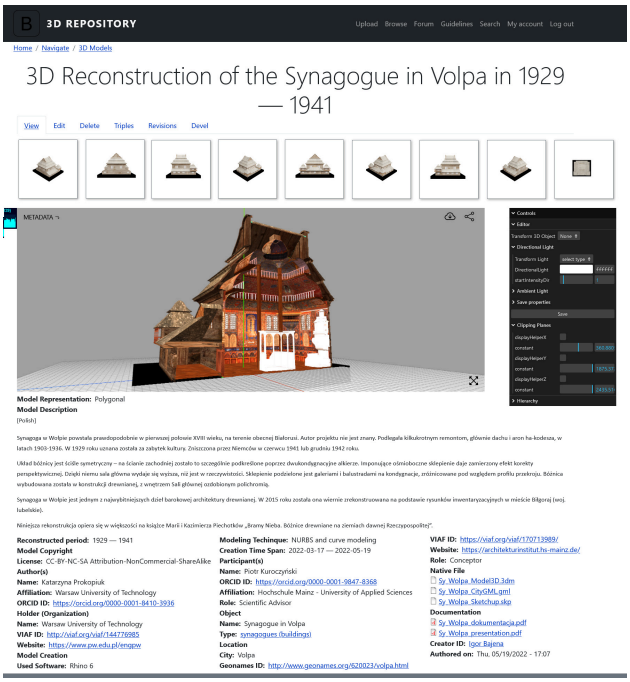


Fig.3
Rendered entity in the 3D Repository with visualized 3D model in the DFG 3D-Viewer (©2022, Hochschule Mainz).

Further research

The next stage of the project is the implementation of the developed modular DFG 3D-Viewer in various academic institutions' repositories, which will be realized in the next two years. The final solution will be available as minimal-effort plugin (set of scripts) for any environment that supports JavaScript, PHP and Python. The datasets from decentralised library repositories will be indexed and displayed in centralized browser web service. As a result, users are provided with a uniform interface for viewing digitised media. The project serves also for further fundamental research conducted by two PhDs in work in the topic of the scientific validation of published 3D reconstruction data and also visualization of the uncertainty on the published 3D models.

Bibliography

Champion, E. and Rahaman, H. (2020). Survey of 3D digital heritage repositories and platforms, *Virtual Archaeology Review,* **11(23)**:1.

https://www.cidoc-crm.org/ (accessed 09 December 2021)

Cieslik, E.(2020). *3D Digitization in Cultural Heritage Institutions Guidebook*. Baltimore: Dr. Samuel D. Harris National Museum of Dentistry.
https://www.dfg.de/en/research_funding/principles_dfg_funding/good_scientific_practice/index.html (accessed 11 April 2022)
http://dfg-viewer.de/en/ (accessed 10 December 2021)
Dworak, D., Kuroczyński, P. (2016) Virtual Reconstruction 3.0 – New Approach of Web-based Visualisation and Documentation of Lost Cultural Heritage. *Proceedings of 6th International Conference EuroMed*, Cyprus: Springer International Publishing LNCS Series, pp. 292–306.
https://www.go-fair.org/fair-principles/ (accessed 10 December .2021)
Fernie, K. et al. (2020).*3D content in Europeana task force*, Hague: Europeana Network Association.
https://kompakkt.de/home (accessed on 10 December 2021).
Kuroczyński, P.(2017). Virtual Research Environment for Digital 3D Reconstructions: Standards, Thresholds and Prospects. In: Frischer, B., Guidi, G., Börner, W., (Hg.) *Cultural Heritage and New Technologies 2016 Proceedings*, *Studies in Digital Heritage,* Open Access Journal, Vol. 1, No. 2, pp. 456–476.
Kuroczyński, P., Bajena, I., Große, P., Jara, K., Wnęk K.(2022) Digital Reconstruction of the New Synagogue in Breslau: New Approaches to Object-Oriented Research. In Niebling, F., Münster, S. (eds.), *Proceedings of the Conference on Research and Education in Urban History in the Age of Digital Libraries & Digital Encounters with Cultural Heritage*, Springer, January 2022.
Messemer, H. (2016) The Beginnings of Digital Visualisation of Historical Architecture in the Academic Field. In: Hoppe, S. and Breitling, S. (eds.), V *irtual Palaces, Part II. Lost Palaces and their Afterlife. Virtual Reconstruction between Science and Media*, München: PALATIUM, pp. 21-54.
Münster, S. (2019) Digital Cultural Heritage as Scholarly Field – Topics, Researchers and Perspectives from a bibliometric point of view In: *Journal of Computing and Cultural Heritage* **12(3)**: 22-49.
https://www.new-synagogue-breslau-3d.hs-mainz.de (accessed on 08 December 2021)
https://www.patrimonium.net (accessed on 10 December 2021)
https://sketchfab.com/ (accessed on 19 April 2022)
http://wiss-ki.eu/ (accessed on 09 December 2021)

Representing uncertainty and cultural bias with Semantic Web technologies

Baroncini, Sofia
sofia.baroncini4@unibo.it
Digital Humanities Advanced Research Centre, University of Bologna

Daquino, Marilena
marilena.daquino2@unibo.it
Digital Humanities Advanced Research Centre, University of Bologna

Pasqual, Valentina
valentina.pasqual2@unibo.it
Digital Humanities Advanced Research Centre, University of Bologna

Tomasi, Francesca
francesca.tomasi@unibo.it
Digital Humanities Advanced Research Centre, University of Bologna

Vitali, Fabio
fabio.vitali@unibo.it
Digital Humanities Advanced Research Centre, University of Bologna

Disagreements on scholarly topics are often the result of different levels of expertise, cultural-dependent viewpoints and methodologies (Eco, 1976; Ginzburg, 1978), as well as geographical and temporal constraints, due e.g. to scholars' provenance or temporal changes in interpreting reality. For example, classifying modern Chinese calligraphy (CMC) artworks is challenging (Iezzi, 2015). For instance, the series of paintings "Da wo miao mo" (1994-) by Zhang Dawo (张大我) [1] (Iezzi, 2014) has been categorised by Gordon Barrass as "oriental abstract expressionism". Wang Nanming categorised it as "abstract expressionism of calligraphic characteristics", stressing on its calligraphic component - cfr. also (Iezzi, 2013-4; Xia Kejun, 2015). Despite not being alternative statements, these reflect differences rooted in scholars' backgrounds - and the scholars' identity is recognized as an important element to understand classifications of CMC (Iezzi, 2015).

Back in 1939, Erwin Panofsky argued that background and experience of the observer can affect even rather simple tasks such as the identification of objects and events

represented in the painting (Panofsky, 1939). Panofsky mentions the baby depicted in van der Weyden's Three Magi [2], who is modernly understood as fluctuating, since he presents attributes traditionally assigned to apparitions (such as being in perspective and in mid-air with no support). Instead, miniatures with Byzantine influences, such as the "Gospels of Otto III" [3] present an irreal empty space around the city of Nain, which does not imply the use of perspective nor that the city is fluctuating - being only an abstraction for decorative purposes. Nowadays, Semantic Web technologies are widely used to formally represent interpretative complexities. However, it has been argued that ontologies carry cultural biases at the schema level (Janowicz, 2018), and hardly integrate different viewpoints in the ontological representation of reality. Thesauri like ICONCLASS (Couprie, 1978), or reference models like CIDOC-CRM (Doerr et al., 2007) do not allow to assign cultural constraints to concepts like the usage of perspective (Baroncini et al., 2021).

Moreover, the semantics associated with RDF and related representation strategies (e.g. n-ary relations, reifications, named graphs (Noy and Rector, 2006; Carroll et al., 2005)) is ambiguous. Statements with different degrees of certainty (whether these are undisputed, currently disputed or settled) are equally represented as assertions, despite their truth value with respect to the dataset is varying (Barabucci et al., 2021). Therefore, when describing the interpretations of Dawo's work as being either connected with "expressionism" or "calligraphy" we are actually asserting both statements without being able to characterize their truth value. Even when specifying provenance and attributions, statements potentially biased by scholar's background (e.g.: {:Da-wo-miao-mo :style :abstract-expressionism-calligraphic} and {:Da-wo-miao-mo :style :oriental-abstract-expressionism}) are both asserted and coexist at the same time in the same knowledge space.

The unclear semantics associated with graph data has some impractical drawbacks. First, a reasoner may interpret competing statements as either being the same or concurring: human intervention would be needed to disambiguate statements as alternative and culturally-dependent interpretations that can only be accepted within a given context. A machine-understandable strategy is needed to express without asserting statements whose truth value depends on the context.

Second, ontologies are needed to annotate uncertainty and truth of statements. Since ontologies are representative of specific cultures, people with diverse backgrounds would struggle to reuse the same terminology, therefore abstaining from expressing diverging information (reticence) or flattening, reducing or coercing their interpretations in a way that conforms to the semantics of the given ontology. Moreover, several models to represent provenance, uncertainty, and truth exist (a recent survey is Sikos and Philp, 2020), making the extraction of information on contexts and uncertainty cumbersome and time-consuming. An ontology-independent solution to represent uncertainty is needed to prevent information loss and to simplify retrieval of data and context information.

While several ontology-independent solutions have been proposed, their efficacy in representing the truth value of statements is limited and unsatisfactory (Barabucci et al., 2021). For instance, named graphs (Carroll et al., 2005) have been widely used to separate assertions from provenance. However, there is no consensus on their semantics, and there are up to eight different model-theoretic semantics to choose from with extremely different takes on the assertiveness of their content (Arndt and Van Woensel, 2019).

In this work, we compare several strategies to represent uncertainty in the Semantic Web. We highlight limits of unclear semantics and demonstrate that common situations in humanistic discourse cannot be unambiguously represented by them. Then we propose an approach to express conjectural statements without asserting them. Conjectures are an extension to RDF 1.1 that by design represents named graphs whose truth value is unknown, regardless of which of the eight semantics for named graphs is chosen (Rolfini, 2021). Through conjectures it is possible to faithfully represent hypotheses, competing or contradictory claims, points of view we agree or disagree with, and even absurdities (Barabucci et al., 2021).

For example, statements on Dawo's series would be represented as follows (in Trig syntax):

```
CONJECTURE :attribution1 { :Da-wo-miao-mo :style :oriental-abstract-expressionism } .
CONJECTURE :attribution2 { :Da-wo-miao-mo :style :abstract-expressionism-calligraphic } .

:attribution1 :wasAttributedTo :gordon-barrass .
:attribution2 :wasAttributedTo :wang-nanming .
```

Listing 1.
Conjectures in Trig syntax

In the example, the addition of the prefix CONJECTURE to the graph definition allows one to express the statement without asserting it. When querying for meanings associated with Ce (in SPARQL, SELECT * WHERE { :Da-wo-miao-mo :style ?style }), results would include an empty set of assertions. Instead, when asking for uncertain meanings (in SPARQL, SELECT * WHERE {CONJECTURE ?c {:Da-wo-miao-mo :style ?style}}) the query would return the two conjectures. Data consumers would therefore understand that no final decision has been taken on the topic, and data can be deemed unbiased. Notice that, in the conjecture-based query, no explicit reference to ontology terms was made when looking for uncertain

statements, since this approach is completely ontology-independent. We conclude stressing the importance, when expressing and querying culturally-biased data, to be able to represent not just provenance information on competing claims, but also their independent and possibly incompatible existence, and the need for an ontology-independent way to express their truth values, as made possible through conjectures.

Future developments include the application of Conjectures over a large knowledge base with the aim of testing the model feasibility on a large scale. At the moment, an online converter from Conjectures to plain RDF is available [4].

Bibliography

Arndt, D. and Van Woensel, W. (2019). Towards supporting multiple semantics of named graphs using N3 rules. *13th RuleML+RR 2019 Doctoral Consortium and Rule Challenge, Proceedings*, vol. 2438. CEUR http://hdl.handle.net/1854/LU-8632551.

Barabucci, G., Tomasi, F. and Vitali, F. (2021). Supporting Complexity and Conjectures in Cultural Heritage Descriptions. *104-115* https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2736994.

Baroncini, S., Daquino, M. and Tomasi, F. (2021). Modelling Art Interpretation and Meaning. A Data Model for Describing Iconology and Iconography. *ArXiv:2106.12967 [Cs]* http://arxiv.org/abs/2106.12967.

Carroll, J. J., Bizer, C., Hayes, P. and Stickler, P. (2005). Named graphs, provenance and trust. *Proceedings of the 14th International Conference on World Wide Web*. (WWW '05). New York, NY, USA: Association for Computing Machinery, pp. 613–22 doi: 10.1145/1060745.1060835.

Couprie, L. D. (1978). Iconclass, a device for the iconographical analysis of art objects. *Museum International*, **30**(3–4). Routledge: 194–98 doi: 10.1111/j.1468-0033.1978.tb02136.x.

Doerr, M., Ore, C.-E. and Stead, S. (2007). The CIDOC conceptual reference model: a new standard for knowledge sharing. *Tutorials, Posters, Panels and Industrial Contributions at the 26th International Conference on Conceptual Modeling-Volume 83*. pp. 51–56.

Eco, U. (1976). *Opera Aperta*. Bompiani Milano.

Ginzburg, C. (1978). Spie, Radici di un paradigma scientifico. *Rivista Di Storia Contemporanea*, **7**(1). Loescher Editore.: 1.

Iezzi, A. (2014). LA 'MODERNITÀ' DELLA CALLIGRAFIA. Metamorfosi e influenza della calligrafia cinese all'interno del panorama artistico contemporaneo. Sapienza Università di Roma PhD dissertation https://opac.bncf.firenze.sbn.it/bncf-prod/resource?uri=TD17046444&v=l&dcnr=6.

Iezzi, A. (2015). What is 'Chinese Modern Calligraphy'? An Exploration of the Critical Debate on Modern Calligraphy in Contemporary China. *Journal of Literature and Art Studies*, **5**(3): 206–16 doi: 10.17265/2159-5836/2015.03.007.

Janowicz, K., Yan, B., Regalia, B., Zhu, R. and Mai, G. (2018). Debiasing Knowledge Graphs: Why Female Presidents are not like Female Popes. *International Semantic Web Conference (P&D/Industry/BlueSky)*.

Noy, N., Rector, A., Hayes, P. and Welty, C. (2006). Defining n-ary relations on the semantic web. *W3C Working Group Note*, **12**(4). World Wide Web Consortium Cambridge, MA, USA.

Panofsky, E. (1939). *Studies in Iconography*. New York: Oxford University Press.

Rolfini, A. (2021). Semantics of Conjectures. *ArXiv Preprint ArXiv:2110.08920* doi: 10.48550/ARXIV.2110.08920. https://arxiv.org/abs/2110.08920.

Sikos, L. F. and Philp, D. (2020). Provenance-Aware Knowledge Representation: A Survey of Data Models and Contextualized Knowledge Graphs. *Data Science and Engineering*, **5**(3): 293–316 doi: 10.1007/s41019-020-00118-0.

Notes

1. See Zhang Dawo 张大我, The star of the city, a rock and roll singer, 2011, ink on paper https://www.researchgate.net/profile/AdrianaIezzi/publication/283087106/figure/fig25/AS:668707621720069@1536443729391/ZHANG-Dawo-The-Star-of-the-City-a-Rock-and-Roll-Singer-2011-ink-on-paper-162-cm-x.ppm
2. See https://www.wga.hu/art/w/weyden/rogier/07bladel/3bladel.jpg
3. See https://www.digitale-sammlungen.de/en/view/bsb00096593?page=66
4. See http://conjectures.altervista.org/convert.html

Beyond the Tracks: connecting people, places and stations to re-assess the impact of rail in Victorian Britain

Beelen, Kaspar
kbeelen@turing.ac.uk
The Alan Turing Institute, United Kingdom

McDonough, Katherine
kmcdonough@turing.ac.uk
The Alan Turing Institute, United Kingdom

Lawrence, Jon
j.lawrence3@exeter.ac.uk
University of Exeter

Rhodes, Josh
jrhodes@turing.ac.uk
The Alan Turing Institute, United Kingdom

Wilson, Daniel C.S.
dwilson@turing.ac.uk
The Alan Turing Institute, United Kingdom

In historical research it is rare to have very detailed information that is: (a) about people in the places they lived (b) broad in its geographical and temporal scope and (c) pertains to complementary aspects of lived experience. Historians construct arguments about the past with far less, writing microhistories or grand narratives that are based on much smaller sets of collected evidence. Here, we introduce a method for creating and connecting high-resolution, geolocated historical information, and we outline an approach for the humanistic interpretation of this evidence. Historical studies using big datasets typically interrogate single source types. More rarely, researchers combine two source types to create new insights (e.g., Gregory and Martí-Henneberg, 2010). We combine three significant new, open datasets in a series of 'convergence experiments' which use place as a means of framing new historical questions. We discuss the challenges and opportunities of making different types of datasets interoperable. Like recent work that makes use of spatial information embedded in text, termed 'geospatial semantics' (Gidal and Gavin, 2019) or geographical text analysis (Taylor and Gregory, 2022), we seek to combine structured, spatial data of different forms.

Census

Most previous work with nineteenth-century British census data has aggregated individual-level data at the parish level to chart national occupational and demographic change over time (Shaw-Taylor and Wrigley, 2014). Our approach retains the national scale but vastly increases the resolution we work at by linking c.70% of individuals in 1881, 1891, and 1901 to the streets they lived on (using Ordnance Survey [OS] Open Roads and GB1900 data [Aucott and Southall, 2019]). Though sampling raises its own issues, there remain significant benefits to working at the more precise level of streets.

Maps

Unlike the census, historical series maps have received almost no attention at the collection, rather than individual sheet, level. We treat the rich details found in large-scale OS Maps as a 'visual census' to be examined computationally alongside the population census (Hosseini et al, 2021a). MapReader is a Computer Vision software library we have developed that creates open, reproducible labeled data based on queries of OS maps (Hosseini et al, 2021b). Producing data using our 'patchwork method' allows us to investigate thousands of maps and to predict the presence of buildings and rail infrastructure across Britain. Unlike other railway track datasets, our data is: (a) more complete and (b) richer (because it includes sheet-level metadata about survey and print dates) and so by its nature, this 'railspace' dataset offers a novel measure of how industrialisation impacted the physical and social landscape.

Stations

Although researchers know the location of Britain's c.12,000 railway stations, this information has not previously been in a structured form with rich attributes like company names and opening and closing dates. Our StopsGB dataset can be linked to 'railspace' to distinguish distinctive aspects of the railway system: the total footprint of railspace v. its specifically passenger-facing structures (Coll Ardanuy et al, 2021).

Convergence

The power of our approach lies in combining the datasets through the prism of place, and being able to do so whilst varying the degrees of geographic precision. In this presentation, we use this approach to investigate not only the much-discussed network-amenity aspects of rail (Bogart et al, 2022), but also the locally negative impacts ('disamenity'). We investigate this ambient or environmental effect in relation to a broader conception of railspace using MapReader patches which, uniquely, capture the wider footprint of the infrastructure beyond the railway track and stations typically vectorised into points and lines.

Triangulating three datasets allows us to investigate the social effects of rail as a key facet of industrialisation. We quantify the spatial relationships between railway infrastructure, street-level socio-demographic data (e.g. the percentage of residents working in particular

economic sectors, or of households with servants), and proximity/access to railway stations (e.g. as a means of access to work). In addition, by calculating the density of 'railspace' per street, we categorise individuals based on their proximity to substantial railway infrastructure (see fig.1). These metrics suggest how rail could operate as an 'amenity' (e.g. a station with little accompanying rail infrastructure) and/or a 'disamenity' in different communities. For instance, we find contrasting proximities to stations among wealthy households (professional/finance workers; multiple servant-keeping; low room occupancy) depending on whether they were in urban or rural areas.

Structuring historically-rich digital sources so that they can be integrated in novel and flexible ways promises to open new perspectives on the social history of industrialization and urbanization. Our method allows scholars to shift easily between big-picture macro analysis, and the fine-grained, human-scale exploration of social context.

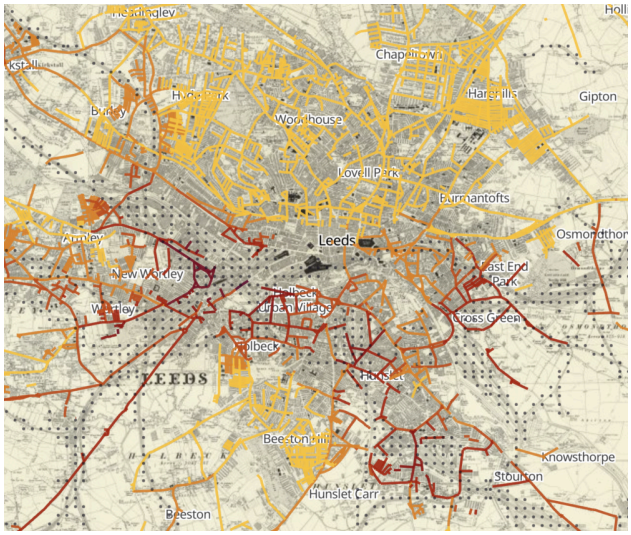


Fig. 1.
Streets in Leeds in 1901 coloured according to their proximity to dense railspace (from yellow to dark red; MapReader railspace patch-centroids shaded grey). Map data courtesy of NLS.

Bibliography

Aucott, P. and Southall, H. (2019). Locating Past Places. *International Journal of Humanities and Arts Computing* 13(1-2): 69-94.

Bogart, D. et al. (2022) Railways, Divergence, and Structural Change. *Journal of Urban Economics* 128: 1-23.

Coll Ardanuy, M. et al. (2021) Station to Station: Linking and Enriching Historical British Railway Data. *CHR 2021: Computational Humanities Research Conference*. Amsterdam: CEUR Workshop Proceedings, pp. 249-265.

Gidal, E. and Gavin, M. (2019). Infrastructural Semantics. *International Journal of Geographical Information Science* 33(12): 2523-2544.

Gregory, I. and Martí-Henneberg, J. (2010) The Railways, Urbanization, and Local Demography. *Social Science History* 34(2): 199-228.

Hosseini, K. et al. (2021a) MapReader. arXiv:2111.15592.

Hosseini, K. et al. (2021b) Maps of a Nation? *Journal of Victorian Culture* 26(2): 284-99.

Shaw-Taylor, L. and Wrigley, E.A. (2014) Occupational Structure and Population Change. In Floud, R., Humphries, J. et al. (eds), *The Cambridge Economic History of Modern Britain*. Cambridge, pp. 53-88.

Taylor, J. and Gregory, I. (2022) *Deep Mapping the Literary Lake District: A Geographical Text Analysis*. Bucknell University Press: Lewisburg, PA.

Machine Learning May Be the Future, but Can It Be the Past? What Machine Learning Systems May Mean for the Historical Concept of Provenance

Benito-Santos, Alejandro
abenito@usal.es
VisUSAL Research Group, Universidad de Salamanca, Spain

Doran, Michelle
doranm1@tcd.ie
Trinity Long Room Hub Arts & Humanities Research Institute, University of Dublin Trinity College

Edmond, Jennifer
edmondj@tcd.ie
Trinity Long Room Hub Arts & Humanities Research Institute, University of Dublin Trinity College

Therón, Roberto
theron@usal.es
VisUSAL Research Group, Universidad de Salamanca, Spain

With a few limited exceptions (Blanke et al., 2020), the application of machine learning (ML) within the historical research process maintains a strong human-in-the-loop element that limits the extent of its proliferation. Part of the reason for this is surely the omnipresence of certain kinds of uncertainty in historical research, which the traditions of historiography have developed powerful (albeit analogue) tools to manage. As Myles Lavan recently suggested, the persistence of these longstanding methods may not be 'a mistaken belief that uncertainty about the past is qualitatively different from that faced by other disciplines,' (Lavan, 2019) however. Instead, we propose that ML/AI methods challenge one of the most fundamental and foundational elements of historical research, namely provenance, in ways that are not simple to resolve or document. In historical research, provenance typically refers to the record of where an object, collection, or dataset has come from and the places and 'experiences' (additions, transformations, deletions, etc.) it has had since its original documentation. The entry of ML methods into DH might be changing the limits and implications of this definition.

For such methods to be meaningfully applied to historical research, provenance needs to be reconsidered, modelled from multiple perspectives, and documented differently from the current standards in computer science. Specifically, data transformations that are no longer performed by human actors but by autonomous or semi-autonomous computational systems need to be captured to enable provenance management. Conversely, historians' reliance on provenance requires that we (a) agree upon a shared definition of data provenance and (b) ensure that ML systems designed for use in this specific context maintain legibility. Such a negotiation between research fields will require more than the current research in explainable AI promises to deliver, making the computational provenance not only be reconstructed but also comprehensible in the multidisciplinary space of DH.

The research project "PROgressive VIsual DEcision-Making in Digital Humanities" (PROVIDEDH, 2017-2021) has contributed to this requirement by proposing a Visual Analytics (VA) approach to representing and managing uncertainty in DH research and demonstrating how a better communication of human or machine-induced uncertainty can enhance the user experience for humanities scholars using ML models. Among other outputs, the project developed an HCI-inspired uncertainty taxonomy (see Figure 1) differentiating between two main types of uncertainty: human-made and technology-made, which correspond to aleatoric (irreducible) and epistemic (reducible) uncertainty as per previous works in the literature (Edmond, 2019; Therón Sánchez et al., 2019; Simon et al., 2018) (Edmond, 2019; Therón Sánchez et al., 2019; Simon, 2017).

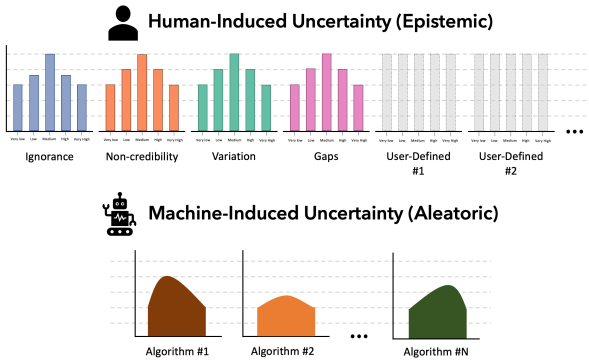


Figure 1:
Proposed uncertainty model. Top: human-induced uncertainty with four predefined categories that map to the epistemic categories previously introduced by Fisher and others. Users can add more categories on a per-project basis if required. Bottom: Machine-induced uncertainty showing the results of applying N different algorithms to the data.

The first uncertainty category, technology-induced, can be mapped to aleatoric uncertainty (well-defined objects in Fisher's (Fisher, 1999) taxonomy) and results from applying computational algorithms to the data, which often give their results with a variable degree of bounded uncertainty (e.g., topic models). For this reason, this type of uncertainty is better represented as a continuous probability distribution. In addition, this representation allows a better understanding of speculative runs of a given algorithm and enhances the what-if analysis process. For example, a researcher could parametrise an algorithm with a fixed set of inputs and launch it several times, obtaining a range of mean values and deviations encoded in a probability distribution function (PDF), which, if correctly displayed, would allow her to get an idea of how the algorithm behaves. Analogously, the algorithm could be parametrised with a variable set of inputs created by the user running the computation or by other researchers. This operation mode would answer the questions of "what happens if I run the algorithm n times using my assumptions?" or "what happens if I run the algorithm n times using another person's assumptions?" As in the case of running the algorithm with the same parameters many times, the results of multiple runs with different parameters could also be summarised in a continuous PDF, allowing the desired kind of what-if analysis. We argue this kind of insights are highly valuable, specifically in the case of probabilistic algorithms, such as topic models or word embeddings, and whose results – and thus, interpretations– can vary significantly between different runs (Alexander and Gleicher, 2016).

The other category, human-induced uncertainty, arises from 1) direct interpretations of the raw data (which in turn may be based on others' previous interpretations and

grounded expert knowledge of the user), 2) interpretations of computational analyses performed on the data, or 3) most likely, a combination of the two. Human actors report this category on a 5-point Likert scale, which is thus best modelled as a discrete PDF. The relationships of dependency between the categories in our taxonomy are bidirectional and self-recurring since, for example, input parameters and data — and therefore the results — are derived from a user's previous interpretations of textual data and related machine- or human-generated annotations. In turn, these interpretations must necessarily be built upon previous insight obtained by the same or other users who apply computational techniques to the data. This creates a temporal belief network (Druzdzel and Simon, 1993; Pearl and Mackenzie, 2018) (see Figure 2) in which the actors' perspectives are fixated on the different versions of a dataset.

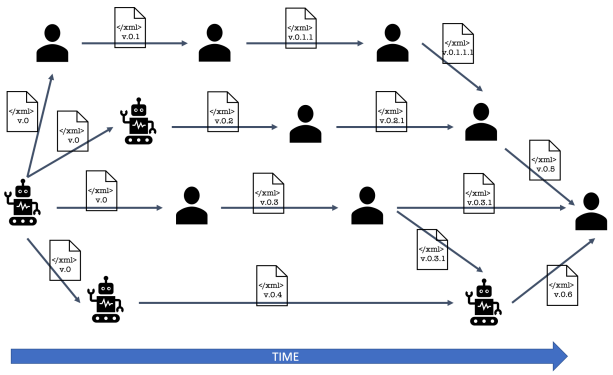


Figure 2:
A Bayesian Belief Network formed by different interactions of a machine and human actors with the data. Each of these interactions produces a new version of the data, which may, in turn, be used as an input by another actor to create a more recent version.

Our taxonomy was evaluated in different user studies (Benito-Santos et al., 2021), and it can be used by other researchers in a digital research platform (https://providedh.ehum.psnc.pl/). Although this kind of encoding may still feel foreign to many researchers trained in the traditions of historical methods, it will only be through this kind of convergence between the affordances and constraints of ML on the one said, and the values and tolerances of historical research on the other, that we will be able to see more widespread integration of ML into historical research workflows.

Bibliography

Alexander, E. and Gleicher, M. (2016). Task-Driven Comparison of Topic Models. *IEEE Transactions on Visualization and Computer Graphics*, **22**(1): 320–29 doi:10.1109/TVCG.2015.2467618.

Benito-Santos, A., Doran, M., Rocha, A., Wandl-Vogt, E., Edmond, J. and Therón, R. (2021). Evaluating a Taxonomy of Textual Uncertainty for Collaborative Visualisation in the Digital Humanities. *Information*, **12**(11). Multidisciplinary Digital Publishing Institute: 436 doi:10.3390/info12110436.

Blanke, T., Bryant, M. and Hedges, M. (2020). Understanding memories of the Holocaust—A new approach to neural networks in the digital humanities. *Digital Scholarship in the Humanities*, **35**(1): 17–33 doi:10.1093/llc/fqy082.

Druzdzel, M. J. and Simon, H. A. (1993). Causality in Bayesian Belief Networks. In Heckerman, D. and Mamdani, A. (eds), *Uncertainty in Artificial Intelligence*. Morgan Kaufmann, pp. 3–11 doi:10.1016/B978-1-4832-1451-1.50005-6. http://www.sciencedirect.com/science/article/pii/B9781483214511500056 (accessed 2 December 2020).

Edmond, J. (2019). Strategies and Recommendations for the Management of Uncertainty in Research Tools and Environments for Digital History. *Informatics*, **6**(3): 36 doi:10.3390/informatics6030036.

Fisher, P. F. (1999). Models of uncertainty in spatial data. *Geographical Information Systems*, **1**: 191–205.

Lavan, M. (2019). Epistemic Uncertainty, Subjective Probability, and Ancient History. *The Journal of Interdisciplinary History*, **50**(1): 91–111 doi:10.1162/jinh_a_01377.

Pearl, J. and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. 1st ed. USA: Basic Books, Inc.

Simon, C., Weber, P. and Sallak, M. (2018). *Data Uncertainty and Important Measures*. John Wiley & Sons.

Therón Sánchez, R., Benito-Santos, A., Santamaría Vicente, R. S. and Losada Gómez, A. (2019). Towards an Uncertainty-Aware Visualization in the Digital Humanities. *Informatics*, **6**(3): 31 doi:10.3390/informatics6030031.

Processing tangles in the *Frankenstein Variorum*

Beshero-Bondar, Elisa Eileen
eeb4@psu.edu

Penn State Erie, The Behrend College, United States of America

Borgia, Mia

mqb5995@psu.edu
Penn State Erie, The Behrend College, United States of America

Chan, Jacqueline

jkc5782@psu.edu
Penn State Erie, The Behrend College, United States of America

Viglianti, Raffaele

rviglian@umd.edu
Maryland Institute for Technology in the Humanities, University of Maryland, United States of America

Computer-aided collation is like a power loom that inevitably tangles up threads caught in the machinery. Automating a tedious process magnifies the complexity of error-correction, calling for new tooling to help us smooth the weaving process. For the DH 2022 conference we seek to share our efforts in the *Frankenstein Variorum* project (hereafter referred to as *FV*) to automate corrections to machine-assisted collation and thereby to refine our collation pre-processing and post-processing algorithms.

FV began during the recent 1818-2018 bicentennial celebrating the first publication of Mary Shelley's novel and exists now as a partial working prototype. We are constructing a digital variorum edition that highlights alterations to the novel *Frankenstein* over five key moments from its first drafting in 1816 to its author's final revisions by 1831. Source "threads" for the *FV* collation "weave" include two well established digital editions: the Pennsylvania Electronic Edition (*PAEE*) an early hypertext edition produced at the University of Pennsylvania in the mid 1990s by Stuart Curran and Jack Lynch, and the Shelley-Godwin Archive's edition of the manuscript notebooks (*S-GA*) published in 2013 by the University of Maryland. That interface relies on a backend processing pipeline that involves the software collateX in collating textual data, including markup, from each edition. To finalize our work requires iterative efforts to refine our collation and post-processing algorithms.

The Gothenburg Model, conceptualized by the developers of collateX and Juxta in 2009 (see https://collatex.net/doc/) , organizes a series of distinct and iterative stages in a workflow for automated collation. These stages involve tokenizing and normalizing the texts to be collated, determining at what points the texts align and diverge, and visualizing the results of collation. Our current efforts on the *FV* involve post-processing the software-generated collation data. We seek to automate the identification of common patterns of misalignment and to produce a more accurate rendering of collation units mapped to the S-GA edition.

At ADHO 2022, we wish to discuss these two efforts with *1. improving collation alignment*, and *2. improving our visualization of the S-GA edition* within our variorum.

1. Improving collation alignment: We are applying XSLT to seek out patterns of "spurious alignment" generated by our collation software, collateX. The collation algorithm tends to err optimistically, seeking alignment of completely divergent passages on single words like "and" or "the." One solution to this is to remove these words entirely during the pre-processing stage, but we rejected this approach because we consider the changes from "and" to "or", or "the" to "an" to be significant. Since we generate collation output as critical apparatus markup in TEI-conformant XML, we are opting to locate patterns of divergence using XSLT as a post-processing stage.

2. Improving our visualization of the S-GA edition: Representing the *S-GA* edition accurately is a challenge because we needed to *re-sequence* its encoding to prepare it for tokenization and normalization in our collation algorithm. In the *S-GA*'s TEI markup, marginalia in the manuscript notebook pages were encoded at the ends of each page file, and they were given attributes that indicate their insertion points in the running flow of the text. It was necessary to re-sequence the order of text on the page to move the marginalia from the end of each file to its insertion point so that we could prepare a continuous sequence of text—the thread of the *S-GA*—to compare with the threads of the other four editions. Resequencing the *S-GA* meant following a clearly signaled trail of ids and pointers in the original encoding. While it would indeed be convenient to display the re-sequenced TEI in our edition viewer, we seek instead to *map* our collation data back onto the original source document by pulling the source document's code into our reading interface. We seek to apply the information we learned from the collation to point back to specific passages in the source document, identifying them by line and string position in the original files in order to pull those particular passages into our interface viewer. This is an ambitious challenge involving stand-off pointers that require counting characters in the source document for precise identification of a variant passage in the *S-GA*'s original encoding.

We need to improve our stand-off pointing mechanism to the *S-GA* edition. Our interface needs to pinpoint in the original *S-GA* files the precise location of the variant

passages identified by the collation process. To improve this, we are revisiting the process of re-sequencing the document in the first place. We are retracing our steps in the early XSLT written for the project to set clearer markers to identify the locations of marginalia passages in the original *S-GA* files. Those markers need to be delivered to the collation output XML, to assist in calculating the XPath and string locations of variant passages in the source S -*GA*.

At ADHO 2022 we will share our efforts in both of these areas, in the hope of encouraging lively discussion in the text scholarly community. If we are to smooth the tangled webs of collation, perhaps we need to be able to follow our own complex algorithms backwards to the threading of the machine.

Who's In and Who's Out: 10 Years On

Bleeker, Elli
elli.bleeker@huygens.knaw.nl
Huygens Institute for the History of the Netherlands – Royal Netherlands Academy of Arts and Sciences

Beelen, Kaspar
kasparvonbeelen@gmail.com
Alan Turing Institute

Chambers, Sally
Sally.Chambers@ugent.be
Ghent University

Koolen, Marijn
marijn.koolen@di.huc.knaw.nl
Huygens Institute for the History of the Netherlands – Royal Netherlands Academy of Arts and Sciences

Melga-Estrada, Liliana
lilianamelgar@runbox.com
0

Van Zundert, Joris J.
joris.van.zundert@huygens.knaw.nl
Huygens Institute for the History of the Netherlands – Royal Netherlands Academy of Arts and Sciences

At the 2011 Convention of the Modern Language Association Stephen Ramsay stirred up a controversy that had been smoldering within the Digital Humanities (DH) community for a while. In a lightning talk Ramsay provocatively asserted that digital humanists should be able to code or else they were "out". Ramsay's statement met considerable criticism (Ramsay, 2013a, 2013b). Ten years on it would probably not provoke as strong a reaction as Ramsay witnessed in 2011. However, it has also not become more clear to what extent programming skills are – or should be – a requirement for practitioners in the field of DH (cf. O'Sullivan, James, Jakacki, and Galvin 2015; Callaway et al. 2020; Van Zundert, Antonijević, and Andrews 2020).

This paper reports on a large-scale survey into the role of code in the digital humanities. The survey used an online questionnaire to investigate what the target community understands by 'code literacy', its relevance for the field of DH, and how it can be promoted among students and researchers. Answering these questions allows us to suggest, first, a shared definition and vocabulary for talking about code literacy grounded in the community's own conceptualizations and norms. It also serves to make recommendations for situating code literacy sustainably in (digital) humanities curricula.

The question of code does not merely concern the availability of practical skills that might be convenient to handle, for instance, digital texts. Rather the matter goes straight to that wretched question at the heart of the community: "What is Digital Humanities?" To code or not to code presents a dichotomy that is strongly coupled with epistemological choices, for which matters whether 5 or 95 percent of a community of practice thinks the ability to create code is an essential analytical faculty or merely a possibly handy tool.

There is no lack of opinions and discussions on the topic, but most are based on anecdotal evidence and personal experience. Some analytical work has been done (cf. Callaway 2020; Van Zundert, Antonijević, and Andrews 2020) but a thorough, comprehensive investigation that both quantitatively and qualitatively surveys the digital humanities landscape with regard to the question remains wanting. For that reason we decided to carry out an investigation in great depth and breadth.

A comprehensive literature review (Webster and Webster 1985; Hockey 1986; Dobberstein 1993; Tafazoli, Parra, and Abril 2017, Tannenbaum 1987; Vee 2013; Montfort 2015; Melgar, Wigham, and Koolen 2019; Earhart et al. 2016; Spante et al. 2018; Potter 2010; Piotrowski and Fafinski 2020; Vee 2017; and so forth) served to establish a conceptual and analytical framework around the issue of 'code literacy'. Our quantitatively-oriented approach consisted, given the complex nature of the topic, of a mixed-method (Timans et al., 2019) online survey that was designed and tested in several reflective iterations over the space of seven months.

The survey gathered demographic information about the respondents, information on their career levels and

phases, self-estimations of code skills, pedagogical context, teaching needs, and provided the ability to submit a definition for 'code literacy'. The survey was globally announced through most well known DH channels. It yielded 399 completed responses in 2.5 months, which provided us with a solid basis to correlate the scores on the various variables with characteristics of given code literacy definitions.

Closed questions have been statistically analyzed to establish significant correlation between variables. For the open question on the definition of 'code literacy' we followed an inductive approach to coding the results. Overlapping samples representing more than two-thirds of all answers were subject to open manually coding to establish inter annotator agreement and a shared set of codes. These codes were then applied to the full data, after which axial coding resulted in a hierarchy of aspects that captures all code literacy aspects shared importantly between many definitions. A methodological overview and first results will be published in *DH Benelux Journal* (Bleeker et al., accepted for publication).

Our initial analysis focused on representativeness, demographics, and their relation to views on code literacy. Results indicate that 93% of a representative sample of the DH community finds code literacy 'at least somewhat important' to 'crucial' while 7% of respondents found code literacy not important at all. Obviously distribution varies with academic background, career level and phase, and self-estimated competence, but overall a majority consensus is clear, even assuming a considerable self-selection effect within the survey's audience makeup.

The fine grained data allows us to infer many more observations and details. In our presentation we will further unpack these results. We will expound what skills and epistemological aspects crucially relate to code literacy according to the respondents. We detail differences in the interpretations and role of code literacy across humanities disciplines. The data also allow us to list the distribution of teaching needs as to skills and techniques, and how current levels of code proficiency were attained.

One of our findings is that 'context is crucial'. That is: the code literacy skills most valued by the most code proficient scholars tend to become developed in the context of research where there is a clear methodological motivation to do so. What also becomes clear from the data is that most code literate scholars, no matter the (self-estimated) level of code literacy, are dissatisfied about their coding proficiency. This provides further evidence that there is a discrepancy between the specific code literacy requirement in DH methodology and the abilities (e.g., within university curricula) to acquire these skills in a concrete humanities context.

Our observations allow us to propose ways for DH scholars and students to develop their code literacy skills, and to provide evidence-based suggestions for the incorporation of code literacy training in curricula. We will disclose results from our current research, notably an ontology of skills based on the practices described by participants through the survey. The ontology thus provides an evidence based means of discussing code literacy, its aspects, and related skills. We hope it will be productive as a resource for planning elements of code literacy training to be embedded in curricula.

Bibliography

Bleeker, E., Beelen, K., Chambers, S., Koolen, M., Melgar-Estrada, L. and Van Zundert, J. (submitted). Persistence, self-doubt, and curiosity: Surveying code literacy in Digital Humanities. *DH Benelux Journal*.

Callaway, E., Turner, J., Stone, H. and Halstrom, A. (2020). The Push and Pull of Digital Humanities: Topic Modeling the 'What is digital humanities?' Genre. *DHQ: Digital Humanities Quarterly*, **14**(1): 450 http://www.digitalhumanities.org/dhq/vol/14/1/000450/000450.html (accessed 6 December 2021).

Dobberstein, M. (1993). Computer literacy for the rest of us. *Computers and the Humanities*, **27**(5): 429–33 doi: 10.1007/BF01829393. https://doi.org/10.1007/BF01829393 (accessed 6 December 2021).

Earhart, A., Kirschenbaum, M., Nowviskie, B., Harris, K., Zafrin, V., Murray, P. J. and Allington, D. (2016). Doing DH vs. Theorizing DH *ACH: Digital Humanities Questions & Answers* https://dhanswers.ach.org/topic/doing-dh-v-theorizing-dh/ (accessed 6 December 2021).

Hockey, S. (1986). Susan Hockey. Workshop on Teaching Computers and the Humanities Courses. *Literary and Linguistic Computing*, **1**(4): 228–229 doi: doi.org/10.1093/llc/1.4.228. https://doi.org/10.1093/llc/1.4.228 (accessed 6 December 2021).

Melgar-Estrada, L., Wigham, M. and Koolen, M. (2019). Programming humanists - What is the role of coding literacy in DH and why does it matter?. *DH Benelux 2019*. Liège: Université de Liège, p. 25 http://2019.dhbenelux.org/wp-content/uploads/sites/13/2019/08/DH_Benelux_2019_paper_25.pdf (accessed 15 September 2019).

Montfort, N. (2015). Exploratory Programming in Digital Humanities Pedagogy and Research. *A New Companion to Digital Humanities*. John Wiley & Sons, Ltd, pp. 98–109 doi: https://doi.org/10.1002/9781118680605.ch7. https://onlinelibrary.wiley.com/doi/

abs/10.1002/9781118680605.ch7 (accessed 6 December 2021).

O'Sullivan, J., Jakacki, D. and Galvin, M. (2015). Programming in the Digital Humanities. *Digital Scholarship in the Humanities*, **30**(suppl_1): i142–47 doi: 10.1093/llc/fqv042. https://doi.org/10.1093/llc/fqv042 (accessed 20 April 2022).

Piotrowski, M. and Fafinski, M. (2020). Nothing New Under the Sun? Computational Humanities and the Methodology of History. *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)*, vol. 2723. Amsterdam: CEUR Workshop Proceedings, pp. 171–181 doi: urn:nbn:de:0074-2723-3. http://ceur-ws.org/Vol-2723/short16.pdf (accessed 5 October 2021).

Potter, W. J. (2010). The State of Media Literacy. *Journal of Broadcasting & Electronic Media*, **54**(4). Routledge: 675–696 doi: 10.1080/08838151.2011.521462. https://doi.org/10.1080/08838151.2011.521462 (accessed 5 October 2021).

Ramsay, S. (2013a). On building. In Vanhoutte, E., Nyhan, J. and Terras, M. (eds), *Defining Digital Humanities: A Reader*. Surrey: Ashgate, pp. 243–46 https://web.archive.org/web/20131205051042/http://stephenramsay.us/text/2011/01/11/on-building/ (accessed 6 December 2021).

Ramsay, S. (2013b). Who's In and Who's Out. In Vanhoutte, E., Nyhan, J. and Terras, M. (eds), *Defining Digital Humanities: A Reader*. Surrey: Ashgate, pp. 239–42 https://web.archive.org/web/20121015012254/http: //stephenramsay.us/text/2011/01/08/whos-in-and-whos-out.html (accessed 6 December 2021).

Spante, M., Hashemi, S. S., Lundin, M. and Algers, A. (2018). Digital competence and digital literacy in higher education research: Systematic review of concept use. (Ed.) Wang, S. *Cogent Education*, **5**(1). Cogent OA: 1519143 doi: 10.1080/2331186X.2018.1519143. https://doi.org/10.1080/2331186X.2018.1519143 (accessed 5 October 2021).

Tafazoli, D., Parra, M. E. G. and Abril, C. A. H. (2017). Computer literacy: Sine qua non for digital age of language learning & teaching. *Theory and Practice in Language Studies*, **7**(9). Academy Publication Co., Ltd.: 716 doi: 10.17507/tpls.0709.02. https://www.researchgate.net/profile/Dara-Tafazoli/project/My-PhD-Thesis-A-cross-cultural-study-on-the-relationship-between-CALL-literacy-and-the-attitudes-of-Spanish-and-Iranian-English-language-students-and-teachers-towards-CALL/attachment/59bc435a4cde26fd91fbe78c/AS:538963140988929@1505510234363/download/1248-4859-1-PB.pdf?context=ProjectUpdatesLog (accessed 5 October 2021).

Tannenbaum, R. S. (1987). How Should We Teach Computing to Humanists?. *Computers and the Humanities*, **21**(4). Springer: 217–25 http://www.jstor.org/stable/30207392 (accessed 6 December 2021).

Timans, R., Wouters, P. and Heilbron, J. (2019). Mixed methods research: what it is and what it could be. *Theory and Society*, **48**(2): 193–216 doi: 10.1007/s11186-019-09345-5. https://doi.org/10.1007/s11186-019-09345-5 (accessed 6 December 2021).

Van Zundert, J. J., Antonijević, S. and Andrews, T. L. (2020). 'BlackBoxes' andTrue Colour — A Rhetoric of Scholarly Code. In Edmond, J. (ed), *Digital Technology and the Practices of Humanities Research*. Cambridge, UK: Open Book Publishers, pp. 123–62 https://www.openbookpublishers.com/product/1108 (accessed 17 February 2020).

Vee, A. (2013). Understanding Computer Programming as a Literacy. *LiCS*, **1**(2): 42–64 https://licsjournal.org/index.php/LiCS/article/view/794 (accessed 30 September 2021).

Vee, A. (2017). *Coding Literacy: How Computer Programming Is Changing Writing*. (Software Studies). Cambridge: The MIT Press.

Webster, S. W. and Webster, L. S. (1985). Computer Literacy or Competency?. *Teacher Education Quarterly*, **12**(2). Caddo Gap Press: 1–7 http://www.jstor.org/stable/23474573 (accessed 6 December 2021).

Exploring Lexical Diversities

Blombach, Andreas
andreas.blombach@fau.de
University of Erlangen-Nürnberg, Germany

Evert, Stephanie
stephanie.evert@fau.de
University of Erlangen-Nürnberg, Germany

Jannidis, Fotis
fotis.jannidis@uni-wuerzburg.de
University of Würzburg, Germany

Pielström, Steffen
pielstroem@biozentrum.uni-wuerzburg.de
University of Würzburg, Germany

Konle, Leonard
leonard.konle@uni-wuerzburg.de
University of Würzburg, Germany

Proisl, Thomas

thomas.proisl@fau.de
University of Erlangen-Nürnberg, Germany

Introduction

The assumption that literary texts are particularly complex is one of the most important premises of work in literary studies (for example Koschorke 2016, Nan Da 2019). This complexity can be perceived on many different levels, with lexical diversity being one of many determining factors. Different disciplines have proposed different measures over time, but only recently some attempts have been made to consolidate research findings into a comprehensive overview (for example Jarvis 2013; Tweedie/Baayen 1998). Here, we propose a multi-dimensional model of lexical complexity. We provide a definition for each dimension and suggest a best-practice operationalization for most. These operationalizations are validated by comparing a collection of texts for adult readers with a collection of comparable texts aimed at children. Finally, we illustrate the usefulness of our approach in application to literary texts. Though we work with German texts, previous work on variability with different languages including Chinese and Japanese has shown that these measures are not language specific (Pielström et al. in preparation).

Corpora

The validation corpora (Weiß & Meurers 2018) contain German non-fiction text from the educational magazine "Geo" (www.geo.de), a publication conceptually comparable to the "National Geographic", and its offshoot for children called "Geolino". For literary texts, we compare highbrow novels(161 works, approx. 17 mio. tokens) with "dime novels" (1167 works in six different genres, approx. 40 mio. tokens), both under copyright. Dime novels are a type of fiction mass-produced in long-lasting series and sold in kiosks rather than book stores.

Aspects of complexity and measurement

Quantifying diversity is no trivial task. As Jarvis (2013b) points out, existing measures of lexical diversity often lack an underlying construct definition and intuitive concepts of diversity vary. Jarvis proposes six dimensions to properly define the construct: variability, volume (which we do not consider separately), evenness, rarity, dispersion, and disparity. Additionally, we look at innovation, surprise, and density.

Variability

The most intuitive indicator of lexical diversity is the variability of the words used in a text. The most widely known measure is the type-token ratio (TTR).

TTR depends systematically on sample size. Among the solutions proposed for this problem, standardized TTRs (STTR) calculated from fixed-length text chunks provide a practical and intuitive solution (Fig. 1).

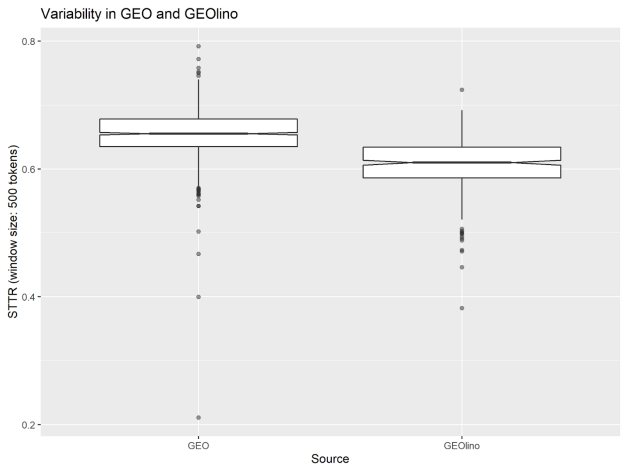


Figure 1:
STTR in GEO and GEOlino

Rarity

A text containing many rare words will generally be perceived as more difficult and more complex than a text with a higher proportion of very common words. We use a simple approach to model rarity. For each text, we compute the proportion of content words not included in the 5,000 most frequent content words from a large web corpus that covers many different registers, the DECOW16BX (Fig. 2, Schäfer and Bildhauer 2012, Schäfer 2015).

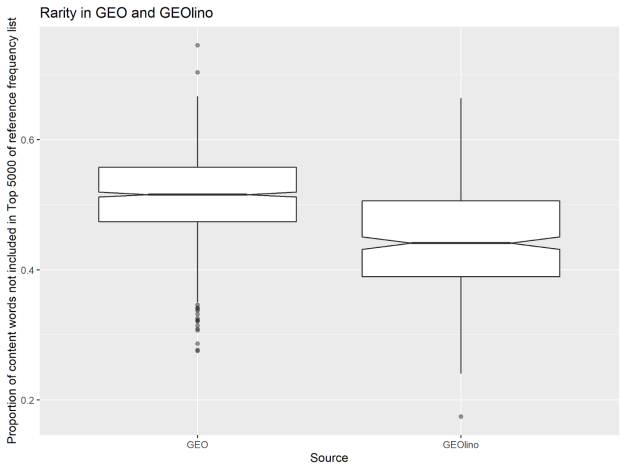


Figure 2:
Rarity in GEO and GEOlino

Dispersion

According to Jarvis (2013b), the perceived lexical diversity is higher if the occurrences of a particular type are more dispersed, whereas a more clustered pattern produces an impression of redundancy. To measure this effect, we again use a window-based approach (Fig. 3). Inside a window, we calculate a dispersion score based on the Gini coefficient (Gini 1912) for each type and use the arithmetic mean of this score over all types with a frequency greater than one as dispersion measure for the whole text (see Blombach et al. in preparation for a detailed description).

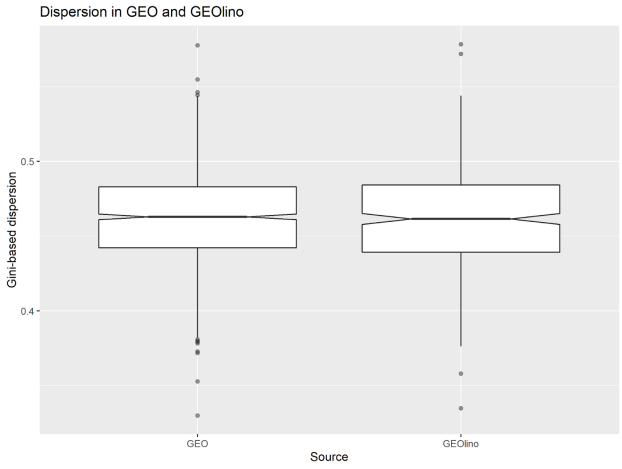


Figure 3:
Dispersion in GEO and GEOlino

Disparity

Lexical disparity follows the intuition that repetition also shows in the occurrence of similarwords on a semantic level. To measure global disparity, a document is segmented and a vector is then generated for each segment by averaging over the vectors of the content words. The disparity of a segment is then calculated from the pairwise euclidean distance of all its segments. The document's disparity is the mean over all its segment disparities (Fig. 4).

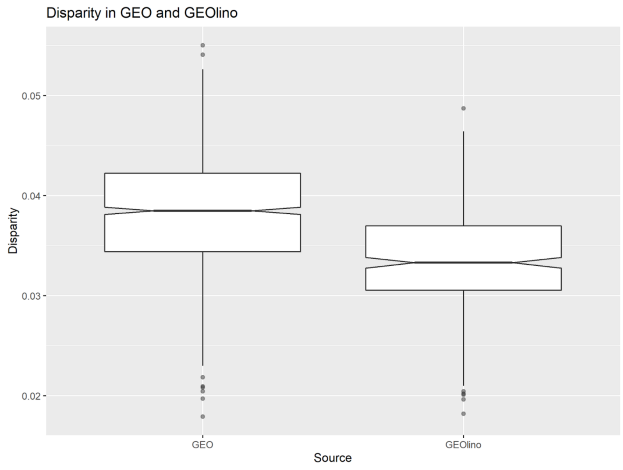


Figure 4:
Disparity in GEO and GEOlino

Density

A text containing a higher proportion of content words can be considered denser and therefore more complex (Fig. 5).

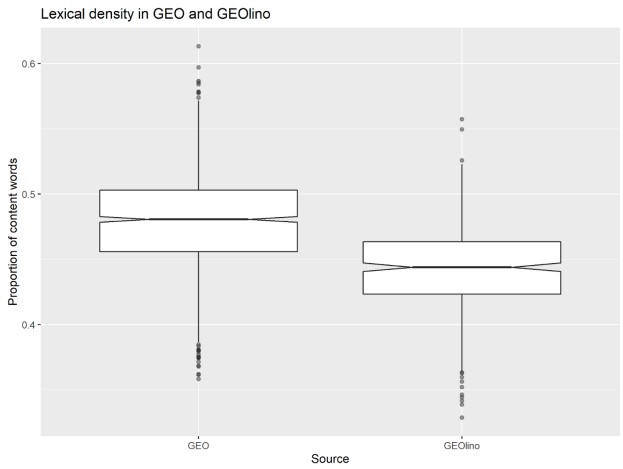


Figure 5:

Density in GEO and GEOlino

Tools

Most of the measures suggested here (variability, rarity, dispersion, and density) are implemented in our textcomplexity toolboxthat contains additional complexity measures as well.

We have also created an interactive “Shiny”appwhich allows users to visually explore our data, including correlations between different measures and the influence of parameters such as window size, case sensitivity and the inclusion or exclusion of punctuation.

Application to Literature

Fig. 6 shows the measures of lexical complexity applied to six genres of dime novels and a set of highbrow novels. Counter to our expectations, science fiction and fantasy equal or even surpass the highbrow novels in some respects (disparity, density, dispersion and rarity). We assume that we have different forms of lexical complexity at work here: In science fiction and fantasy, a noun-heavy prose is depicting new worlds with new words. In high literature on the other hand, high variability shows the influence of a stylistic ideal which aims to avoid repetition and show elegance. There might be a difference in the scope which authors control for complexity, for example variability. We found less repetition in small windows in genre texts, whereas variability in highbrow literature increases with window size.

Fig. 7 shows that genre similarities can be perceived immediately using this kind of representation. A multi-dimensional model of lexical complexity allows a clearer understanding of genre differences.

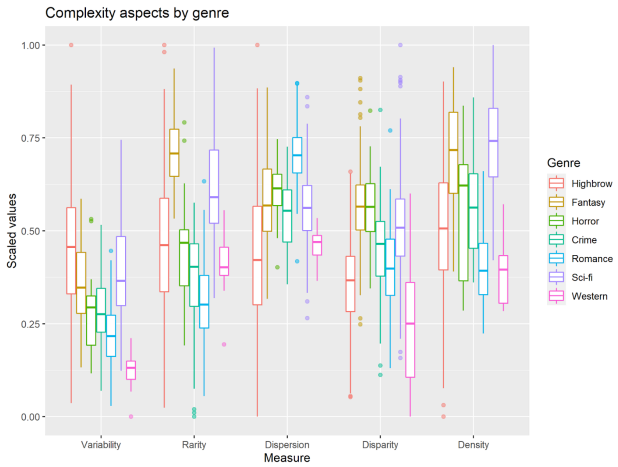


Figure 6:

Diversity per aspect. All dimensions have been scaled to values between 0 and 1

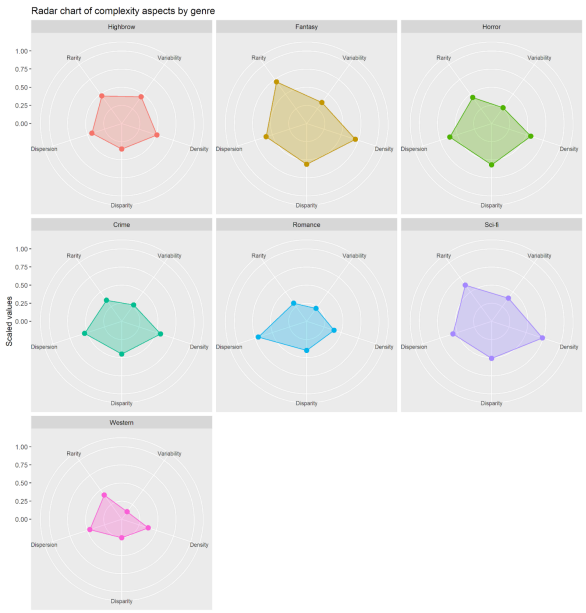


Figure 7:
Radarplots, highlighting the similarities between genres

Bibliography

Blombach, A., Evert, S., Jannidis, F., Konle, L., Pielström, S. and Proisl, T. (in preparation): Lexical Complexity in Texts. A Multidimensional Model.

Da, N. Z. (2019): The computational case against computational literary studies. In: *Critical Inquiry*, 45(3), p. 601–639.

Falk, I., Bernhard, D. and Gerard. C. (2014): From Non Word to New Word: Automatically Identifying Neologisms in French Newspapers. In: *Proceedings of LREC 2014*.

Gini, C. (1912): *Variabilità e Mutuabilità. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche*. C. Cuppini, Bologna.

Jarvis, S. (2013a): Capturing the Diversity in Lexical Diversity. In: *Language Learning* 63 (1), p. 87–106.

Jarvis, S. (2013b): Defining and Measuring Lexical Diversity. In: Jarvis, Scott / Daller, Michael (Eds.): *Vocabulary Knowledge. Human Ratings and Automated Measures*. Amsterdam: John Benjamins. (= Studies in Bilingualism 47)

Klosa, A. and Lungen, H. (2018): New German Words: Detection and Description. In: *Proceedings of the XVIII EURALEX,* p. 559–569. Ljubljani.

Koschorke, A. (2016): *Komplexität und Einfachheit*. p. 1–10. Stuttgart.

Ney, H., Essen, U. and Kneser, R. (1994): On structuring probabilistic dependences in stochastic language modelling. In: *Computer Speech & Language*, Volume 8, Issue 1, p. 1-38.

Pielou, E.C. (1966): The measurement of diversity in different types of biological collections. In: *Journal of theoretical biology*. 13: p. 131–144. doi:10.1016/0022-5193(66)90013-0

Pielström, S., Hodošček, B., Calvo Tello, J., Henny-Krahmer, U., Jannidis, F., Schöch, C., Du, K., Uesaka, A. and Tabata, T. (in preparation): Measuring Lexical Diversity of Literary Texts.

Schäfer, R. (2015): Processing and Querying Large Web Corpora with the COW14 Architecture. In: *Proceedings of Challenges in the Management of Large Corpora* (CMLC-3) (IDS publication server), p. 28–34.

Schäfer, R. and Bildhauer, F. (2012): Building Large Corpora from the Web Using a New Efficient Tool Chain. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, p. 486–493.

Weiß, Z. and Meurers, D. (2018): Modeling the readability of German targeting adults and children: An empirically broad analysis and its cross-corpus validation. In: *Proceedings of the 27th International Conference on Computational Linguistics*, p. 303–317, Santa Fe, New Mexico, USA.

Support and Relationship Patterns in Endometriosis Narratives

Bologna, Federica
fb265@cornell.edu
Cornell University, USA

Thalken, Rosamond
ret85@cornell.edu
Cornell University, USA

Mimno, David
mimno@cornell.edu
Cornell University, USA

Wilkens, Matthew
wilkens@cornell.edu
Cornell University, USA

Introduction

In online support communities, people with endometriosis share experiences of navigating interpersonal relationships while living with a chronic illness. These stories suggest patterns in illness narratives that are unique to relationships with doctors, friends, and family and the kind of support patients receive. Scholarship and literature in the medical humanities have emphasized the importance of understanding relationships in the experience of pain. Therefore, we study the framing of relationships in online endometriosis medical narratives, by designing a classifier to identify the connections between characters, subject matter, and intent in patients' stories. Resulting predictions suggest narrative arcs in the interpersonal endometriosis experience, both in and across posts, as well as societal changes in the awareness of endometriosis. We hope our methodology will be implemented in future research on illness narratives and contribute to the advancement of medical humanities.

Related Literature

Language is the main tool available to the sick to make their pain visible to others, receive care, as well as spark collective action (Scarry, 1987). Studying illness narratives is thus essential to improve medical care, especially for marginalized communities whose experiences of pain may not be adequately acknowledged. Contemporary medical knowledge and practices are tied to the disenfranchisement of certain peoples, as they are informed with Enlightment androcentric and heteronormative notions (Lupton, 2012; Ussher, 2003; Laqueur, 2003).

Women's health, and endometriosis specifically, are complex subjects to consider in the study of illness narratives. Female pain is often dismissed, as the female body's distinctive qualities are considered deviations, and hysteria discourse continues to characterize contemporary medical literature (Young et al., 2019). When treating genital conditions, women's ability to fulfill their reproductive duties is given priority over their need for relief or pleasure (Farrell and Cacchioni, 2012; Scully and Bart, 1973; Ussher, 2003; Lupton, 2012). Consequently, women who suffer from sexual pain are faced with disbelief from their physicians, who perceive them as "difficult" for not accepting the medical system's failure at treating them (Feldhaus-Dahir, 2011; Jones, 2015).

In addition, relationships hold a special role in how the marginalized experience illness. Female lung-cancer patients are faced with pity, fear, or judgment by friends and family, who either give the patient up for dead, worry about being contaminated, or blame them for their condition

(Sontag, 1978; Tomalia, 2014). The fear of contagion is a recurring element of AIDS narratives as well. Here, however, managing the disease and spreading awareness becomes a collective effort in which friends and partners take action. Indeed, many AIDS stories are recounted by people close to the patient (Wojnarowicz, 1991; Guibert et al., 2020). Lastly, for some patients, relationships are their only chance to be heard and to receive proper care, and thus have the power to free or imprison them (Keller, 2010; Gilman, 2018)

Data

Online endometriosis support communities exist on many platforms and in many forms. We focus on Reddit because it has two thriving endometriosis support communities, the subreddits r/Endo and r/endometriosis. Posts from r/Endo span from January 23rd 2012 to September 30th 2021, and r/endometriosis from November 23rd 2014 to September 30th 2021.

	r/Endo	r/endometriosis
Total number of posts	20,479	10,793
Median number of tokens per post	135	132
Unique authors	9,074	6,475

We merge the data from r/endometriosis and r/Endo based on substantial similarities in post content and comparable statistics of the subreddits (table 1).

Supervised Classification

To study the framing of common roles in endometriosis narratives, we design a supervised classification task to identify character tropes, and the post's expressed intent. Character tropes are a set of likely relationship types that often occur in an endometriosis narrative, including family, partner, doctor, and endometriosis support community.

We then look at the different roles a person tends to assume in online health communities, like seeking or providing support (Yang et al., 2019). We use McDowell and Antoniak's (2020) labels for intent, including providing emotional support, seeking emotional support, providing informational support, seeking informational support, providing experiences, seeking experiences, plus an additional label, venting. This intent category turns the focus onto the post's author, to find the role they assume as a member of the community. After labeling posts from the combined r/Endo and r/endometriosis datasets, we train and test a series of binary DistilBERT classification models for both labels.

To explore themes in endometriosis illness narratives, we perform topic modeling at the level of the paragraph. After experimenting with LDA parameters and manually validating several different runnings of the algorithm, we find that the most suited model for our purposes is obtained when setting the number of topics to 25 and removing the 10 most frequent words in the collection. Among the identified topics, a few are dedicated to expressing empathy and gratitude confirming the importance of the patients' relationship to the community for their well-being.

By comparing topic model distributions with predictions about character tropes and expressed intent, we identify how community members narrativize their personal relationships by considering patterns in predicted labels. Changes in the framing of these relationships demonstrate changes in individual and societal experiences of living with endometriosis. We hope that our relationship models can be used in future humanities research on support communities and illness narratives.

Conclusion

By studying the framing of relationships in endometriosis support communities, we draw attention to the myriad ways endometriosis patients find support through their doctors, partners and family, and online community. We consider how such support might also be lacking, resulting in the patient relying on another relationship for emotional, experiential, and informational support. Highlighting stories about relationships in the endometriosis support community demonstrates the significance of shared experience and community when living with chronic illness.

Bibliography

Farrell, J. and Cacchioni, T. (2012). The Medicalization of Women's Sexual Pain. *Journal of Sex Research*, **49**(4): 328–36 doi:10.1080/00224499.2012.688227.

Feldhaus-Dahir, M. (2011). The causes and prevalence of vestibulodynia: a vulvar pain disorder. *Urologic Nursing*, **31**(1): 51–54.

Gilman, C. P. (2018). *The Yellow Wallpaper*.

Guibert, H., Durbin, A., White, E. and Coverdale, L. (2020). *To the Friend Who Did Not Save My Life*. South Pasadena, CA: Semiotext(e).

Jones, C. E. (2015). Wandering Wombs and 'Female Troubles': The Hysterical Origins, Symptoms, and

Treatments of Endometriosis. *Women's Studies*, **44**(8): 1083–113 doi:10.1080/00497878.2015.1078212.

Keller, H. (2010). *The Story of My Life*. New York: Signet Classics.

Laqueur, T. (2003). *Making Sex: Body and Gender from the Greeks to Freud*. 10. print. Cambridge, Mass.: Harvard University Press.

Lupton, D. (2012). *Medicine as Culture: Illness, Disease and the Body*. 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom: SAGE Publications Ltd doi:10.4135/9781446254530. http://sk.sagepub.com/books/medicine-as-culture-3e (accessed 1 December 2021).

McDowell, L. and Antoniak, M. (2020). Symptoms, Scares, and Misclassifications: Information Sharing Behavior Across Online Birth Control Communities. Paper presented at the Black in AI: Workshop at NeurIP, Virtual.

Scarry, E. (1987). *The Body in Pain: The Making and Unmaking of the World*. New York: Oxford university press.

Scully, D. and Bart, P. (1973). A funny thing happened on the way to the orifice: women in gynecology textbooks. *The American Journal of Sociology*, **78**(4): 1045–50.

Sontag, S. (1978). *Illness as Metaphor*. New York: Farrar, Straus and Giroux.

Tomalia, T. (2014). The 'Why Me' of Cancer *A Lil Lytnin' Strikes Lung Cancer* https://lil-lytnin.blogspot.com/2014/12/the-why-me-of-cancer.html.

Ussher, J. M. (2003). I. Biology as Destiny: The Legacy of Victorian Gynaecology in the 21st Century. *Feminism & Psychology*, **13**(1): 17–22 doi:10.1177/0959353503013001003.

Wojnarowicz, D. (1991). *Close to the Knives: A Memoir of Disintegration*. New York: Vintage Books.

Yang, D., Kraut, R. E., Smith, T., Mayfield, E. and Jurafsky, D. (2019). Seekers, Providers, Welcomers, and Storytellers: Modeling Social Roles in Online Health Communities. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Glasgow Scotland Uk: ACM, pp. 1–14 doi:10.1145/3290605.3300574. https://dl.acm.org/doi/10.1145/3290605.3300574 (accessed 18 October 2021).

Young, K., Fisher, J. and Kirkman, M. (2019). 'Do mad people get endo or does endo make you mad?': Clinicians' discursive constructions of Medicine and women with endometriosis. *Feminism & Psychology*, **29**(3): 337–56 doi:10.1177/0959353518815704.

Distant reading of Russian Soviet diaries (Prozhito Database)

Bonch-Osmolovskaya, Anastasia
abonch@gmail.com
National Research University Higher School of Economics

Vorobieva, Viktoria
victoria.vrbva@gmail.com
National Research University Higher School of Economics

Kriukov, Artem
kryukov.ai39@gmail.com
National Research University Higher School of Economics

Podriadchikova, Maria
mpodr2015@gmail.com
National Research University Higher School of Economics

Introduction

The diary as a genre has always been ambiguous and complex, balancing "between literary and historical writing, between the spontaneity of reportage and reflectiveness of the crafted text, between selfhood and events, between subjectivity and objectivity, between the private and the public" (Langford and West 1999, 8). For a long time, researchers have been struggling to define its specifics, and the historical framework has only recently been replaced by an intertextual approach. It has become realizable mostly due to the works of French structuralist Philippe Lejeune (Lejeune 2009).

Due to its genre ambiguity and comprehensiveness, the diary gains a lot of interest both from researchers and common readers. In Russia, this interest was embodied in the Prozhito project—a large database of intimate papers written by people in the Russian Empire and modern Russia but mostly in the Soviet Union. Created in 2015, *Prozhito* has become a source for reflection on history on both national and personal scale.

The *Prozhito* database offers an opportunity to explore diaries on a large scale, with computable methods. However, opportunities go hand-in-hand with challenges, and from this perspective, a researcher who studies the *Prozhito* database faces certain intricacies:

- The *Prozhito* database is heterogeneous. Unlike printed literary works, not a lot of diaries are well preserved. Unsurprisingly, the diaries written by prominent people or created during dramatic times have more chances to be saved.
- The diary is an intimate and personal writing, and this aspect influences the data as well. Typos and

incomprehensible passages inevitably affect the research process and its outcomes.

Nevertheless, the *Prozhito* database is still an encompassing source of material. From a quantitative perspective, several research questions can be asked:

- How various historical events are reflected in the diaries of that time?
- How do people express their feelings in writing? Are there any patterns that people use to describe their emotions?
- Assuming that diaries are diverse and heterogeneous, is it possible to create a classification that would take this heterogeneity into account?

To answer these questions, we conducted several researches based on Digital Humanities methods—topic modeling, collocation analysis, and classification tasks.

How historical events are reflected in the diaries

We compared Soviet publicist articles with diaries created at the same time in order to define how historical names or events are covered in these different texts.

With topic modeling applied, we found that collocations like “Soviet power” (“sovetskaya vlast’”), “class struggle” (“klassovaya bor’ba”), and “class enemy” (“klassovyi vrag”) were popular both in the articles and diaries throughout the whole period presented in data. Some of them like the phrase “enemy of the people” (“vrag naroda”) saw a dramatic growth during the time of the enormous purges.



Figure 1.
Number of the bigram “enemy of the people” during the selected period.

With distant and close reading methods combined, we could find how people expressed their attitude towards these events in writing. We found a correlation between the records mentioning the Civil (1917-22/23) and the Second World wars (1939-45) in which people compared their severe experience and conditions.

These outcomes were defined owing to analysis of collocations and co-occurrences. We tried a similar approach in answering the second question—whether quantitative methods are viable for analyzing emotions in diaries.

How to analyze emotions expressed in writing

Using the bootstrapping method, 6283 records connected with the love topic were selected. It was challenging to define words and phrases used for expressing feelings of love owing to flexible language structure. Nevertheless, splitting data into groups was fruitful in comparing patterns during different time periods.

We started by comparing the proportion of emotional records depending on the period. Due to the database's heterogeneity, we found two periods with a large proportion of emotional records—at the beginning of the 20th century and during the Second World War.

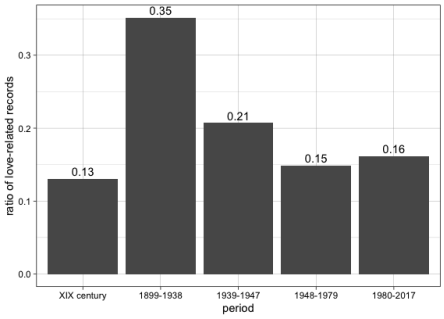


Figure 2.
Emotional records ratio by period.

We selected repeating collocations and analyzed how their frequency changed over time. The postwar period that had not a lot of emotional records has seen a dramatic growth of several bigrams such as “to love passionately” (“strastno lubit’”) and “to love terribly” (“uzhasno lubit’”). The ratio of these bigrams is plummeting during the Second World War when other bigrams like “beloved person” (“lubimyi chelovek”)

rocketed. Being separated, people expressed their sorrow in writings to beloved ones.



Figure 3.
Ratio of the selected bigrams by period.

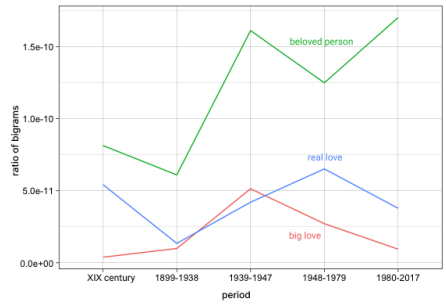


Figure 4.
Ratio of the selected bigrams by period.

The way people describe feelings is changing throughout history. Undoubtedly, shifting in words' meaning makes these explorations challenging. Nevertheless, certain strong patterns can be observed despite these limitations.

How to classify heterogeneous data

Diaries balance between historicity and emotionality, which raises classification issues. We concluded that a proper classification should be based on the parameters apart from genre and author. Instead, we can classify diary records relying on the author's intention to express a certain type of information—a description of a particular life episode, emotion, or even literary text.

Adding such aspects as eventual density and writing style, a preliminary classification can be prepared:

Tag	Object of description	Eventual density	Writing style
NAR	Occurred events: everyday life	High	Informal
WORK	Occurred events: specific activity	High or middle	Mostly formal
EMO	Feelings, emotions and reflections	Middle or low	Informal
EPISODE	Occurred events: everyday life	Low	Informal
LIT	Fictional events	Low	Artistic

After deleting short (less than 500 symbols) texts, we trained a logistic regression model on randomly selected records—documents were represented as vectors using the TfidfVectorizer tool.

At the next stage, in a team of annotators, we manually marked up two datasets twice—one with unbalanced data consisting of 3780 records and a balanced dataset with 2240 records. After annotating, we evaluated the consistency of markup.

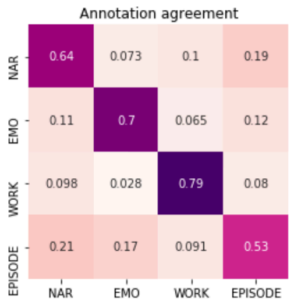


Figure 5.
Consistency of markup

Using built datasets, we tried various linear classifiers such as logistic regression, Naive Bayesian classifier, and Support vector machine. Logistic regression shows better results than other models (f-measure=0.81).



Figure 6.
The error matrix of the basic solution on the balanced dataset.

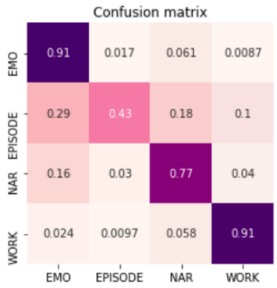


Figure 7.
The error matrix of the basic solution on the unbalanced dataset.

EPISODE and NAR categories were ambiguous. After combining them into one category NAR_EPISODE, the f-measure of the category increased; for a balanced dataset it was 0.78, for an unbalanced one—0.88.



Figure 8.

The error matrix of the basic solution on the balanced dataset.

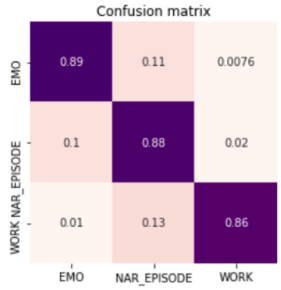


Figure 9.
The error matrix of the basic solution on the unbalanced dataset.

The resulting classification can be variously applied. It becomes possible to analyze what topics are peculiar for authors of a certain age and how individual narrative behavior alters throughout life.

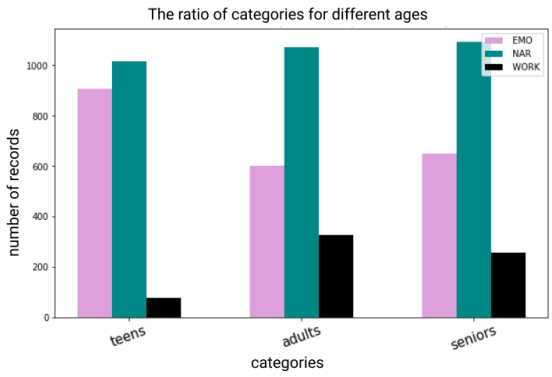


Figure 10.
Ratio of different categories in the diaries of authors of different ages.

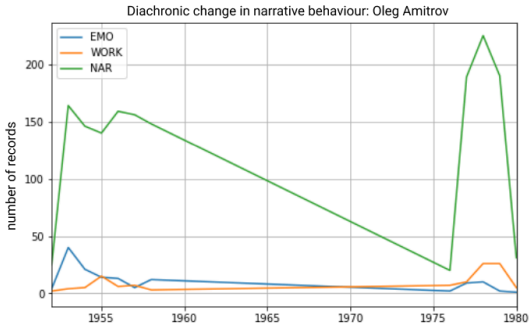


Figure 11.
Oleg Amitrov's narrative behavior—his diaries are among the most thorough in the database.

Conclusion

In our project, we present a framework based on a quantitative method for approaching diaries. We considered both historical and emotional aspects, and these investigations resulted in designing a classification, which, in its turn, is fruitful for new opportunities of computable research in this field.

Bibliography

Langford, Rachael and Russell West. (1999). Introduction: Diaries and Margins. *Marginal Voices, Marginal Forms: Diaries in European Literature and History*. Amsterdam: Rodopi. pp. 6-21.

Lejeune, Philippe. (2009). *On diary*. Honolulu: University of Hawaii Press.

Linguistic Injustice in Multilingual Technologies: arTenTen and esTenTen as case studies

Bordonaba-Plou, David
davidbordonaba@gmail.com
Universidad de Granada, Spain; Universidad de Valparaíso, Chile

Jreis-Navarro, Laila M.
ljreis@unizar.es
Universidad de Zaragoza, Spain; Universidad de Granada, Spain

Establishing English as the *lingua franca* in Academia has contributed to what has become known as "linguistic injustice" (see Van Parijs, 2002; Hyland, 2016; Politzer-Ahles et al., 2016; Soler, 2020). This idea states that second-language learners are at a disadvantage when using the new language. The predominance of English and the difficulties that a poor command of the language can pose have been a relevant concern in Digital Humanities (DH) (see Mahony, 2019, p. 384; Galina, 2014, p. 314). When discussing the consequences and possible solutions to the English-speaking bias in DH, the literature focuses on the following problems: i) the lack of translations of research output (see Galina, 2013); ii) issues of connectivity, for example, the so-called "digital divide" (Galina, 2014, p. 314; Mahony, 2019, p. 385); iii) the unavailability of data sources in languages other than English to quantify DH (Galina, 2014, p. 310); iv) problems with digital standards such as Unicode and TEI (see Fiormonte, 2012, pp. 67-69; Mahony, 2019, p. 374); and v) increasing the number of sources of textual information in languages other than English (see Galina, 2014, p. 314). In this sense, multilingual DH critiques seem to be centered on the deficiencies of non-English-language resources rather than on the level of accuracy of digital analytical tools.

The aim of this work is twofold. Firstly, to distinguish a phenomenon that produces a new type of linguistic injustice, which we label as "the paradox of Anglocentric multilingualism." This paradox arises when a multilingual philosophy is pursued in constructing complex systems of analysis in a digital environment (digital platforms, ontologies). However, these systems imply advantages in the study of English over other languages. The injustice derives from a poor level of precision in the output of technology when analyzing non-English languages. Secondly, we contend that multilingual DH should address the different challenges posed by this paradox. Multilingual DH needs to deal with the deficiencies of tools' performance as well as those of language resources, because this disadvantage makes it difficult for any cross-linguistic study to provide reliable empirical data in (dis)proving linguistic intuitions.

To illustrate some of the potential problems derived from the paradox, this work will detail the difficulties we have faced in a cross-linguistic study on color terms (Bordonaba-Plou and Jreis-Navarro, forthcoming), when using the Arabic corpus arTenTen (Arts et al., 2014) and the Spanish corpus esTenTen (Kilgariff and Renau, 2013) in Sketch Engine. We will study the different performances of the tool in Arabic and Spanish, compared to English, to point out the weaknesses of this tool in a multilingual

arena, making it possible to improve it and enriching the critical and inclusive framework of multilingual DH. Two main issues emerged in our inquiries. Firstly, the lists of collocations provided by the Sketch Word tool are of differing types and usefulness. For example, in Spanish, the tool provides lists like those in English (enTenTen20), i.e., lists based on a functional perspective. However, in Arabic, the researcher has fewer lists available, and those only reflect grammatical categories and the collocation position (left or right). This shortcoming of the tool means it does not provide a complete perspective on the linguistic behavior of the term. Secondly, the analyses conducted by the tool show different degrees of accuracy in Part of Speech (PoS) tagging. In Spanish, the PoS tagger classifies proper nouns as modifiers and verbs. For example, it invents *paular* "to paul" (from the proper name Paula)," or *rodrigar* "to Rodrigo" (from the proper name Rodrigo)." In Arabic, the PoS tagger does not cover cliticization in smart search. Cliticization is the most important phenomena in Arabic morphology and a big challenge in computational analysis (Habash, 2010, pp. 47-50). For example, the tagger reads *bāhit* "pale" as if the prepositional particle proclitic *b+* "with/in" were attached to the non-existent inflected base word *āhit*. These inaccuracies imply that the statistical scores –MI-score (Hunston, 2002, p. 71; Baker, 2006, p. 101), t-score (Hunston, 2002, p. 73), and Log Dice (Gablasova, Brezina and McEnery, 2017, p. 164)– cannot be used with total confidence.

Bibliography

Arts, T., Belinkov, Y, Habash, N., Kilgarriff, A. and Suchomel, V. (2014). arTenTen: Arabic Corpus and Word Sketches. *Journal of King Saud University - Computer and Information Sciences*, **26**(4): 357-371.

Baker P. (2006). *Using Corpora in Discourse Analysis*. Continuum.

Bordonaba-Plou, D., and Jreis-Navarro, L. M. (forthcoming). A cross-linguistic study of color terms in Arabic and Spanish. In Bordonaba-Plou, D. (ed.), *Experimental Philosophy of Language: Perspectives, Methods and Prospects*. Springer.

Fiormonte, D. (2012). Towards a cultural critique of the Digital Humanities. *Historical Social Research*, **37**(3): 59-76.

Gablasova, D., Brezina, V., and McEnery, T. (2017). Collocations in corpus-based language learning research: identifying, comparing, and interpreting the evidence. *Language Learning*, **67**(S1): 155-179.

Galina Russell, I. (2013). Is there anybody out there? Building a global Digital Humanities community. *Humanidades Digitales*, http://humanidadesdigitales.net/blog/2013/07/19/is-there-anybody-out-there-building-a-global-digital-humanities-community/ (accessed 15 August 2021)

Galina Russell, I. (2014). Geographical and linguistic diversity in the Digital Humanities. *Literary and Linguistic Computing*, **29**(3): 307-316.

Habsh, N. Y. (2010). *Introduction to Arabic Natural Language Processing*. Morgan & Claypool.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge University Press.

Hyland, K. (2016). Academic publishing and the myth of linguistic injustice. *Journal of Second Language Writing*, **31**: 58-69.

Kilgarriff, A., and Renau, I. (2013). esTenTen, a vast web corpus of Peninsular and American Spanish. *Procedia - Social and Behavioral Sciences*, **95**: 12-19.

Mahony, S. (2018). Cultural diversity and the Digital Humanities. *Fudan Journal of the Humanities and Social Sciences*, **11**: 371-388.

Van Parijs, P. (2002). Linguistic Justice. *Politics, Philosophy and Economics*, **1**(1): 59-74.

Politzer-Ahles, S., Holliday, J. J., Girolamo, T., Spychalska, M. and Berkson, K. H. (2016). Is linguistic injustice a myth? A response to Hyland (2016). *Journal of Second Language Writing*, **34**: 3-8.

Soler, J. (2020). Linguistic injustice and global English: Some notes from its role in academic publishing. *Nordic Journal of English Studies*, **19**(3): 35-46.

TikTok Cover Dances as Folkloric Practice: Pose Estimation and the Study of Variation in K-Pop Choreography across Short-Form Social Media Videos

Broadwell, Peter
broadwell@stanford.edu
Stanford University, United States of America

Tangherlini, Timothy R.
tango@berkeley.edu
University of California, Berkeley, United States of America

The sharing of short-form videos on social media platforms (TikTok, Instagram and Facebook stories, YouTube Shorts), which grew rapidly in popularity as personal interactions moved online during the COVID-19 pandemic, provides a new opportunity for individual Internet users to participate in the already vibrant, global

online culture surrounding the choreographic aspects of South Korean "idol" (solo or group singer-dancer) K-pop. In addition to the formal dance videos that accompany the release of most new songs and music videos, any Internet user with a smartphone now can record, edit and post their own performance of a brief portion (typically 15-60 seconds) of the song's choreography, accompanied by a publisher-approved snippet of the music. These cover dances usually are centered upon the chorus sections of the songs, which feature the most recognizable choreography and music, and often are prompted by semi-official social media "dance challenge" campaigns. Cover dances that prove popular on a given platform may garner enormous numbers of views, shares, likes/upvotes, comments and new followers (subscribers) for their posters; such recordings also may be included in long-form "highlight reel" compilation videos, which serve as unofficial archives of these otherwise ephemeral posts.

The posting, sharing, and cross-influencing of short-form social media cover dances fits the accepted definition of folklore, digital and otherwise: informal cultural expressive forms circulating over time on and across social networks. Whether the creation and sharing of these videos is motivated by a simple wish to contribute to a perceived community and tradition, or by competition, or by desire for the emotional and monetary spoils of the online attention economy (or some combination of these motivations), the folkloristic frame provides illuminating perspectives on this phenomenon. It contextualizes the short-form cover dances and the choreography in the official videos as bound together in the core dialectic of the folkloric process: the relationship between informal repetitions/renditions enacted by the individual vs. the "tradition," here represented by the indexical demonstration video, but also—and importantly—by the growing number of other individual dancers presenting their own variants of the indexical dance. The degree and manner by which the informal cover dance versions deviate from the indexical form and from each other, such as by incorporating the personal preferences of individual participants or even significant elements of other tradition groups (e.g., other dance styles) highlights how the milieu of globalized K-pop cover dances enacts this central dynamic of the folkloric process.

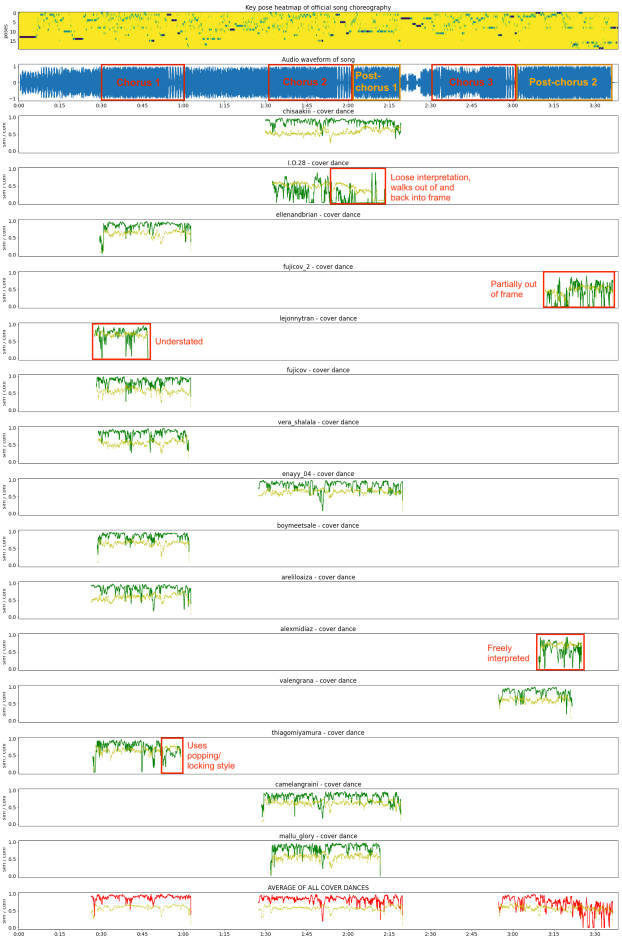


Figure 1.
Pose similarities of 15 short cover dances on TikTok to the original choreography (keypose heatmap on line 1) of BTS's "Permission to Dance"

Neural deep learning-based image processing models that can locate with sufficient accuracy the individual appendages of human poses enable analyses of the interrelationships of dance cover videos at a representative scale. Running such pose estimation tools on dance videos enables both "distant viewing" studies that combine observations from more videos than human observers could reasonably view and annotate, and also "close viewing" analyses that incorporate subtle details of position and motion that would be indiscernible to observers in real time.

For this study, we identified a set of recent high-profile K-pop dance challenges; for each, we ran a pose estimation model optimized for single-figure videos from smartphone cameras, which can interpolate the positions of obscured appendages, on a dozen or so short-form cover dance videos from TikTok as well as on a full-length instructional video demonstrating the song's complete choreography. The model estimates the two-dimensional coordinates of the 17 standard pose keypoints for each frame of a video,

which we normalized within a unit space (L2-norm) to cancel out differences in camera distance and body size. We also augmented each pose with the positions of its 17 keypoints flipped across the horizontal axis so that dances presenting the mirror-image of the choreography —a common occurrence in dance videos—are scored as similar. The similarity between two poses was quantified as the cosine similarity of their normalized keypoint vectors.

Because short-form dance covers for a given song use verbatim excerpts of the music, they can be aligned accurately to the long-form dance demonstration video and to each other by sliding the audio frequency spectrum values of the shorter video across the longer and finding the alignment with the highest cross-correlation score. We then compared the dance poses of the aligned sections by calculating the cosine similarity between any two poses that occurred at the same point in the music. The resulting time series of pose similarity scores enabled us to identify the sections of the original choreography and music that were most commonly "covered" on TikTok, and to detect portions of overlapping choreography that exhibited comparatively greater or lower similarity. Constructing an average of the poses across all covers of a given section also facilitated study of how individual covers (as well as the demonstration video) compare to this emergent consensus form.

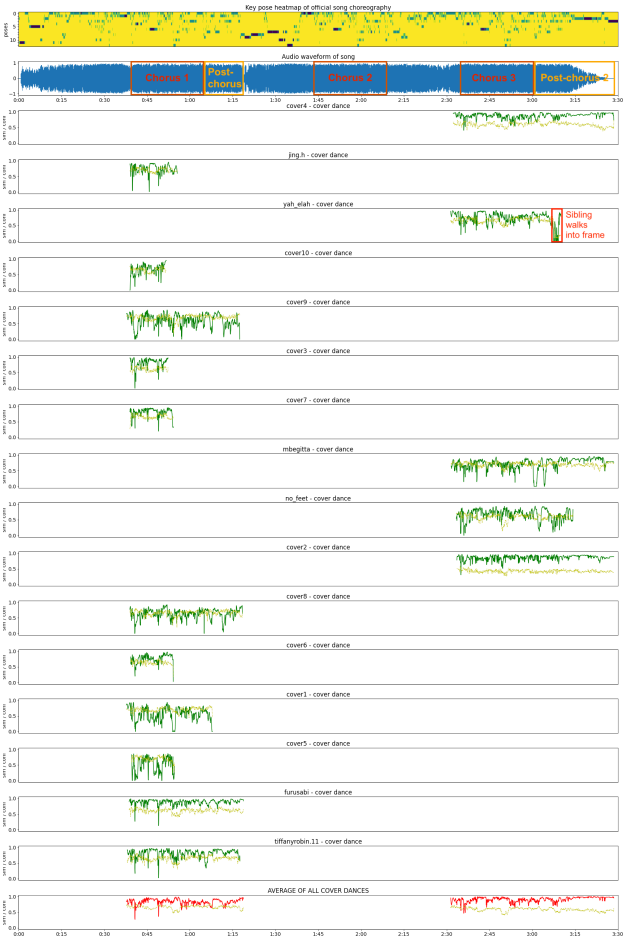


Figure 2.
Plots as in Figure 1 for the dance demonstration video and 16 short cover dances from TikTok for TWICE's "I Can't Stop Me"

Our analytical results (Figures 1 and 2) revealed that most individuated deviations from the indexical and consensus choreography tended to occur at the beginnings and especially the endings of the cover dances, as well as during the most dynamic dance moves (jumping, crouching). Comparing the latter cases to the pose estimation confidence levels indicates that these are not just artifacts of the model "losing track" of keypoints during such moves, but rather, the disruptive nature of the move seems to give the dancers license to deviate from the norm. Similarly, having already established their mimetic proficiency in the earlier sections of a cover dance, performers appear to feel more freedom to introduce personalized variations in the closing portion.

From the perspective of folkloristics, this work is a major contribution: we can interrogate the processes of individual creativity, group predilections, regional variations and the conservative nature of tradition all at once, thus instantiating a macroscopic potential for the study of

popular, informal, yet culturally expressive dance at many scales of analysis.

Bibliography

Abidin, C. (2020). Mapping Internet Celebrity on TikTok: Exploring Attention Economies and Visibility Labours. *Cultural Science Journal* 12(1): 77-103.

Boffone, T. (2021). *Renegades: Digital Dance Cultures from Dubsmash to TikTok*. Oxford University Press.

Broadwell, P. and Tangherlini, T. (2021). Comparative K-Pop Choreography Analysis through Deep-Learning Pose Estimation across a Large Video Corpus. *Digital Humanities Quarterly* 15(1). http://digitalhumanities.org/dhq/vol/15/1/000506/000506.html

Marshall, W. (2019). Social Dance in the Age of (Anti-)Social Media: Fortnite, Online Video, and the Jook at a Virtual Crossroads. *Journal of Popular Music Studies* 31(4): 3-15.

Schellewald, A. (2021). Communicative Forms on TikTok: Perspectives from Digital Ethnography. *International Journal of Communication* 21. https://ijoc.org/index.php/ijoc/article/view/16414

Zulli, D. and Zulli, D. (2020). Extending the Internet Meme: Conceptualizing Technological Mimesis and Imitation Publics on the TikTok Platform. *New Media and Society*. https://doi.org/10.1177/1461444820983603

Tools as Epistemologies in DH? A Corpus-Based Exploration

Burghardt, Manuel
burghardt@informatik.uni-leipzig.de
Leipzig University, Germany

Luhmann, Jan
luhmann@informatik.uni-leipzig.de
Leipzig University, Germany

Niekler, Andreas
aniekler@informatik.uni-leipzig.de
Leipzig University, Germany

Introduction

Invarsson (2021) suggests a digital epistemology that is to be "understood as an attempt to do digital humanities without being committed to digital tools and objects". While it is an intriguing idea to leave digital tools and methods out of the equation in order to discern the true epistemological core of DH, we believe that it is equally possible to argue that the use of specific tools virtually shapes and influences the epistemology of DH, or as Nietzsche (1882) put it: "Unser Schreibzeug arbeitet mit an unseren Gedanken [1]". Today, Nietzsche's observations on writing tools can be easily extended to all kinds of research tools, allowing us to ask questions about the epistemological implications of tools for the digital humanities (see Dalbello, 2011; Drucker, 2002; Ramsay & Rockwell, 2012). As there are manifold, rather diverse tools that are used in DH, there is a certain tradition for tool directories that systematically list and categorize different tools. One of the most popular directories is TAPoR [2], which has steadily evolved and by now includes more than 1,600 tools.

The TAPoR list of tools has been used lately to extract and analyze tools mentioned in DH abstracts (Barbot et al. 2019, Fischer & Moranville, 2020b) and tutorials (Fischer & Moranville, 2020a). The motivation of these analyses is primarily to identify relevant and widely used tools in order to make them sustainably available via infrastructures like the *Social Sciences & Humanities Open Marketplace* [3]. While the previous studies so far have only looked at comparatively small corpora, we suggest to enhance the scope of DH tool studies by using a large corpus of DH journal articles (*Computers and the Humanities*, *Digital Humanities Quarterly*, *Literary and Linguistic Computing*/ *Digital Scholarship in the Humanities*). The corpus comprises 3,737 articles and covers a time span from 1966-2020, which allows for diachronic analyses of tool usage in DH. In addition to using a larger corpus, we also propose an approach to automatically increase the size of the TAPoR tool list.

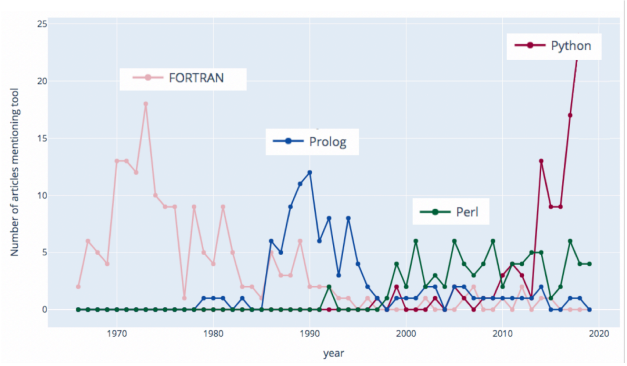
Experiments with the TAPoR tool list

For our experiments we followed the approach described by Barbot et al. 2019 and also used the TaPOR directory to derive queries to find tool occurrences in our corpus. All in all, we found 319 different tools being mentioned throughout the corpus [4]. Figure 1 shows the most frequent 25 tools in the overall corpus.

rank	tool	type	doc_freq	TAPoR URL
1	FORTRAN	programming language	174	https://tapor.ca/tools/4184
2	R / RStudio	data analysis / statistics	119	https://tapor.ca/tools/13481
3	SNOBOL (String Oriented Symbolic Language)	programming language	111	https://tapor.ca/tools/4183
4	Twitter	social network	110	https://tapor.ca/tools/1353
5	Statistical Package for the Social Sciences (SPSS)	data analysis / statistics	95	https://tapor.ca/tools/1288
6	Prolog	programming language	95	https://tapor.ca/tools/494
7	PL/1	programming language	95	https://tapor.ca/tools/2175
8	Oxford Concordance Program (OCP)	text analysis tool	88	https://tapor.ca/tools/417
9	LISP	programming language	81	https://tapor.ca/tools/495
10	Python	programming language	69	https://tapor.ca/tools/1356
11	Perl	programming language	66	https://tapor.ca/tools/1252
12	Excel	data analysis / statistics	65	https://tapor.ca/tools/1485
13	TACT	text analysis tool	55	https://tapor.ca/tools/231
14	JavaScript	programming language	55	https://tapor.ca/tools/831
15	COCOA	text analysis tool	54	https://tapor.ca/tools/222
16	COBOL (Common Business-Oriented Language)	programming language	52	https://tapor.ca/tools/486
17	ALGOL (Algorithmic Language)	programming language	51	https://tapor.ca/tools/487
18	WordCruncher	text analysis tool	44	https://tapor.ca/tools/265
19	SPITBOL (Speedy Implementation of SNOBOL)	programming language	43	https://tapor.ca/tools/491
20	Nota Bene	annotation tool	39	https://tapor.ca/tools/893
21	TUSTEP	text analysis tool	38	https://tapor.ca/tools/201
22	MiniTab	data analysis / statistics	34	https://tapor.ca/tools/1481
23	Collate	text analysis tool (collation)	34	https://tapor.ca/tools/208
24	Pliny	annotation tool	31	https://tapor.ca/tools/343
25	Zotero	reference management tool	31	https://tapor.ca/tools/798

Top 25 tools mentioned in the corpus.

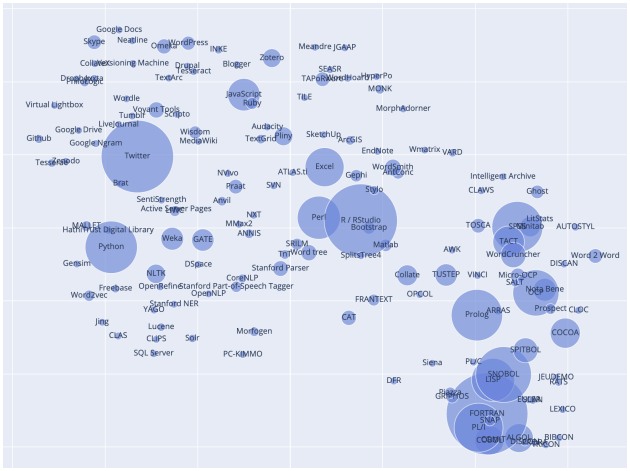
It is noticeable that among the top 25 DH tools we find many tools for text and data analysis, but also a large proportion of high-level programming languages. This is certainly a bias of the early days of humanities computing. Taking a closer look at the diachronic development reveals the natural rise and fall of programming languages, with the steady rise of *Python* in the DH since the beginning of the 2010s being particularly prominent (see Figure 2) [5].



The rise and fall of programming languages in DH.

To provide some more high-level insights, we also did a cooccurrence analysis of the most frequent 150 tools, i.e. tools that are mentioned together in an article (see Figure 3). Such cooccurrence analyses could yield insights into typical tool combinations and more complex workflows, for instance the use of *Brat* to annotate *Twitter* data and the use of *SentiStrength* to perform sentiment analyses of tweets.

All in all, Figures 1-3 show some strong potential to analyze tools to find out more about their epistemological implications of DH. However, the predominance of outdated programming languages and the absence of state-of-the-art tools such as *spaCy* or transformer architectures clearly shows large gaps in the TAPoR list.



UMAP 2D projection of the top 150 most-frequent tools and their nPMI cooccurrence scores (point size indicates numbers of occurrences).

Query expansion experiments by means of tool embeddings

To fill these gaps, we conducted a second experiment in which we used BERT embeddings to expand our list of tools, as proposed by Wevers & Koolen (2020). We created embeddings for the 25 most frequent tools (see Figure 1) and looked for their nearest neighbors, i.e. words that have similar embedding vectors as the 25 most frequent tools (see Figure 4).

	python		rstudio		fortran	
	word	similarity	word	similarity	word	similarity
1	javascript	0.7786302549	rsyntaxtree	0.672282529	mustran	0.779345739
2	java	0.7227259133	nsdstat+	0.6716485016	snobol	0.7478913676
3	python3	0.7176140459	rst	0.664684688	algol	0.7477378491
4	php	0.7055463948	lextutor.ca	0.6579621254	spitbol	0.746969054
5	xml	0.7012426673	langtutor	0.6558568772	fort	0.7423708177
6	html	0.7008138628	aldatu	0.6518317081	snobol4	0.7417163026
7	pascal	0.6852396147	nsdstat	0.6445412307	sgml	0.740959805
8	html5	0.6774953289	destituo	0.640477455	cobol	0.7389697017
9	script	0.673969821	rkdartist	0.6380852832	javascript	0.7368756119
10	programming	0.6711670549	jntolva@artsci.wustl.edu	0.6332783638	programming	0.7327201349

Ten nearest neighbors for the embeddings of "python", "rstudio" and "fortran", ranked by their cosine similarity.

This approach allows us to identify tools that are not listed in the TAPoR directory directly, but that are mentioned in similar article contexts as the TAPoR-listed tools. Obviously, not all the nearest neighbors identified in this way are actual tools, but if we rank the results according to the number of nearest neighborhoods for the top 25 tools, there are indeed many promising results in the higher ranks [6]. To give just one example: *XSLT* (Extensible Stylesheet Language Transformations) has a fairly high score of 16, which means it was in the nearest neighbors of 16 of the 25

most frequent tools from the initial list. In the top ranks we find many other promising tool candidates, such as *RDF*, *SGML*, *mySQL* and also more generic concepts, such as *NLP* and *parsing*, which could be interpreted as tool super categories.

Conclusion and next steps

The experiments by Barbot et al. 2019 and Fischer & Moranville, 2020a/b as well as our follow-up experiments with a larger corpus of texts demonstrate that the empirical analysis of tool mentions in DH publications can be used to discern patterns in the diachronic use of different types of tools. This allows us to explore the effects of tools as rapidly evolving epistemological frameworks in the DH. At the same time, it became clear that the static list of tools as provided by TAPoR has obvious gaps, as the tool landscape is evolving swiftly. We therefore plan to include further directories in follow-up studies, including *ProgrammingHistorian* [7], *forTEXT* [8], *DigiHum* [9], *DMI* (Digital Methods Initiative) [10], *DH Toychest* [11], etc. In this article, we illustrated the benefits of an embeddings-based approach to further expand these static lists of tools. Our next steps will be to extend our corpus to also include articles from neighboring disciplines, such as computational linguistics, computational social sciences, information science and others. We also plan to expand the nearest neighbor search beyond the limit of the 25 most frequent tools and to filter the results list manually, to identify reasonable tools.

Bibliography

Barbot, L., Fischer, F., Moranville, Y. & Pozdniakov, I. (2019). Which DH tools are actually used in research? Published via weltliteratur.net – A Black Market for the Digital Humanities, https://weltliteratur.net/dh-tools-used-in-research/

Bush, V. (1945). As we may think. The Atlantic Monthly, 176(1), 101-108.

Dalbello, M. (2011). A genealogy of digital humanities. Journal of Documentation.

Drucker, J. (2002), “Theory as praxis: the poetics of electronic textuality”, Modernism/Modernity, Vol. 9, November, pp. 683-91.

Fischer, F. & Moranville, Y. (2020a). DH tools mentioned in "The Programming Historian"? Published via weltliteratur.net – A Black Market for the Digital Humanities, https://weltliteratur.net/dh-tools-programming-historian/

Fischer, F. & Moranville, Y. (2020b). Tools mentioned in DH2020 abstracts. ublished via weltliteratur.net – A Black Market for the Digital Humanities, https://weltliteratur.net/tools-mentioned-in-dh2020-abstracts/.

Ingvarsson, J. (2020). Digital Epistemology: An Introduction. In Towards a Digital Epistemology (pp. 1-28). Palgrave Macmillan, Cham.

Nietzsche, F. (1882). Letter 202. An Heinrich Köselitz in Venedig (Typoskript). Nietzsche Source – Digital Critical Edition (eKGWB): http://www.nietzschesource.org/#eKGWB/BVN-1882,202

Ramsay, S., & Rockwell, G. (2012). Developing things: Notes toward an epistemology of building in the digital humanities. Debates in the digital humanities, 75-84.

Wevers, M., & Koolen, M. (2020). Digital begriffsgeschichte: Tracing semantic change using word embeddings. Historical Methods: A Journal of Quantitative and Interdisciplinary History, 53(4), 226-243.

Notes

1. Translation: “Our writing tools [in Nietzsche’s case: his new typewriter] shape our thoughts”.
2. TAPoR: https://tapor.ca/home
3. SSH Open Marketplace: https://marketplace.sshopencloud.eu/
4. For a complete list see https://docs.google.com/spreadsheets/d/1BtDVo_2A6a1cLPQCZ8CriNcSGHz3UuVc2mxTsM-ZCxo/edit?usp=sharing
5. An interactive version of the plots in Figures 2+3 alongside with more plots can be found here: https://bbrause.github.io/tools-in-dh/
6. The nearest neighbors for each of the 25 most frequent tools as well as a ranked overall list is available here: https://docs.google.com/spreadsheets/d/1iipWwyk7wVcaSzzpEq-W5_vjTe2FO-dmjk_IGjMitEc/edit?usp=sharing
7. Programming Historian: https://programminghistorian.org/en/lessons/
8. forTEXT: https://fortext.net/
9. DigiHum: https://digihum.de/tools/
10. DMI: https://wiki.digitalmethods.net/Dmi/ToolDatabase
11. DH Toychest: http://dhresourcesforprojectbuilding.pbworks.com/w/page/69244319/Digita

Evaluation of Multilingual BERT in a Diachronic, Multilingual, and Multi-Genre Corpus of Bibles

Calvo Tello, José
calvotello@sub.uni-goettingen.de
Göttingen State and University Library

De la Rosa, Javier
versae@nb.no
The National Library of Norway

BERT and its Application to Digital Humanities

Transformers (GPT, BERT) have become a central piece in NLP (Vaswani et al., 2017; Alammar, 2018; Tunstall et al., preprint). These language models bring new possibilities for pre-training algorithms with no labelled data, which can then be fine-tuned to new tasks (transfer learning) with fewer labels (few-shot learning). Their linguistic prowess has spurred discussions about their limitations, biases and societal and environmental impact (Bender et al., 2021; Carlini et al., 2021; Underwood, 2021).

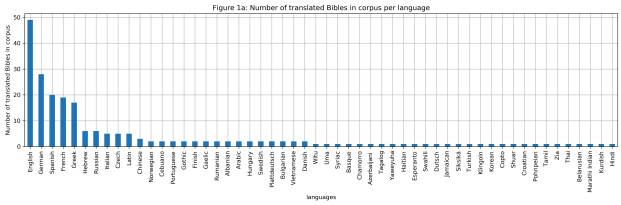
These algorithms have also sparked the interest of the DH community. Specifically, BERT models (Devlin et al., 2019)have been explored mainly for the study of English (Han and Eisenstein, 2019; Sims et al., 2019; Fonteyn, 2020; Underwood, 2021)and German literature (Jannidis and Konle, 2020; Pagel et al., 2021; Ehrmanntraut et al., 2021), or in multilingual settings(de la Rosa, et al., 2021).

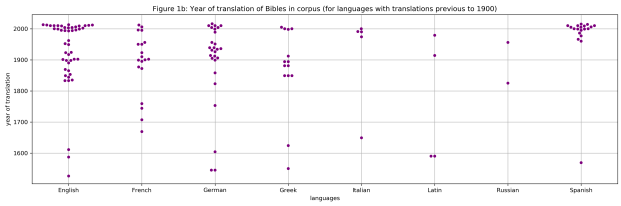
However, its applicability to other Humanities domains remains questionable. Because of the vast amounts of data required for pre-training, these models are usually fed with contemporary text types (web, journalistic, or administrative documents) from resource-rich modern languages. In contrast, the Humanities deal with diverse and heterogeneous datasets, non-standard orthography, historical genres, multilingual datasets, and often from low-resource languages. To assess the performance of multilingual models in this context, we analyze the abilities of a multilingual BERT model (mBERT) pre-trained on Wikipedia for 102 languages (Devlin, 2018)on a multi-genre, diachronic Bible dataset.

Corpus of Bible Translations

Building multilingual corpora usually involves collecting texts produced originally in each analyzed language and period (Odebrecht et al., 2019; Novakova and Siepmann, 2020; Burnard et al., 2021). However, cultural and historical differences hinder the comparability of the results (Schöch et al., preprint). To circumvent this limitation while accounting for low-resource languages, we choose a corpus of translations of Bibles. Bibles as research objects have a long tradition in Corpus Linguistics and Digital Humanities (Radday, 1973; Neumann, 1990; Holmes, 1991; Resnik et al., 1999; Christodouloupoulos and Steedman, 2015; Walczak, 2015; Lee and Yeung, 2016; Calvo Tello, 2020).

The corpus comes from Zefania-XML-bibleand Bible Gateway.It contains 221 translations (11,455 books, e.g. Genesis and Psalms) in 54 languages from all continents, including historical ones such as Latin, Gothic and Syriac, and artificial languages (e.g., Esperanto).Figures 1a-b show the number of translations per language and the historical distribution for languages with translations before 1900.





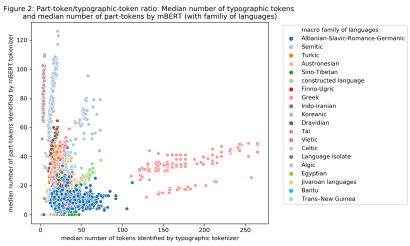
Methods

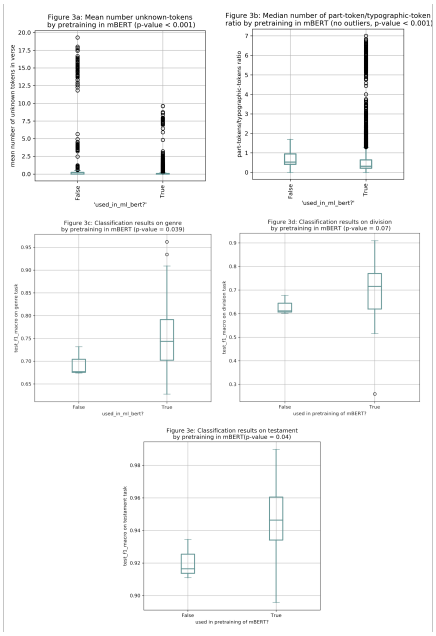
For the analysis, we apply five metrics on two fronts:

1. Tokenization: We apply the default mBERT tokenizer and a simple typographic tokenizer to the Bible dataset. The tokens identified by the mBERT tokenizer can be characters, sub-word units (part-tokens), or whole-word units resembling typographic tokens. For languages using white-space delimited writing systems, the higher the number of whole-word units, the easier to interpret for humans, as the resulting tokenized text

resembles quite closely the original one. To formalize this notion, we calculate the ratio of mBERT part-tokens by typographic-tokens (Figure 2). We also count the number of unknown tokens, i.e. tokens that are not part of the tokenizer vocabulary and cannot be split into its constituents. Low scores in these metrics represent better results.

2. Classification tasks: We apply classification to three annotated categories for each book: genre (e.g., historical, law, letter, Zimmermann, 2010), historical group (e.g., Gospel, Pentateuch), and Testament (Old and New). We create balanced datasets and guarantee each language has at least the same number of Bibles per category. We then split in a training (80%), validation (10%), and test set (10%) and build models for each language as well as combined ones for each century (except for translations pre-1500). We assess the performance of each model using the F1-macro metric (the higher, the better).

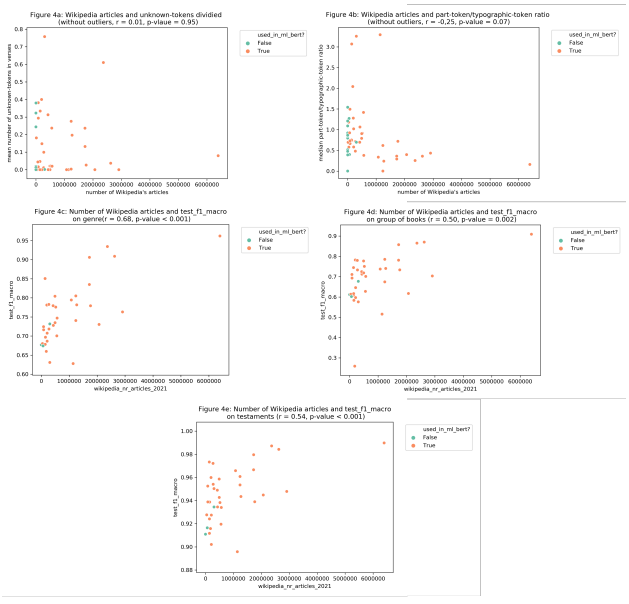




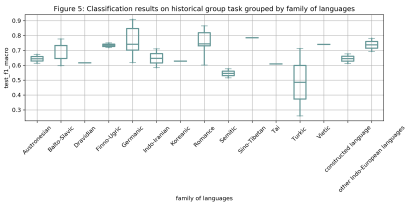
Hypothesis Testing

We use the five tokenization and classification metrics to test several hypotheses on the capabilities of mBERT. First, we hypothesize that languages in the mBERT pre-train dataset should obtain better results in the five metrics. Thus, we expect languages in the pre-train dataset to obtain low values in the tokenization metrics, and high F1 classification scores. To test this hypothesis, we group Bibles by whether their languages were part of the pre-training of mBERT and run Welch's t-tests on the tokenization (Figures 3a-b) and classification (Figures 3c-e) metrics. This hypothesis is supported by three of the five criteria: the number of unknown-tokens (Figure 3a) and the classification results on genre and Testament (Figure 3c and 3e).

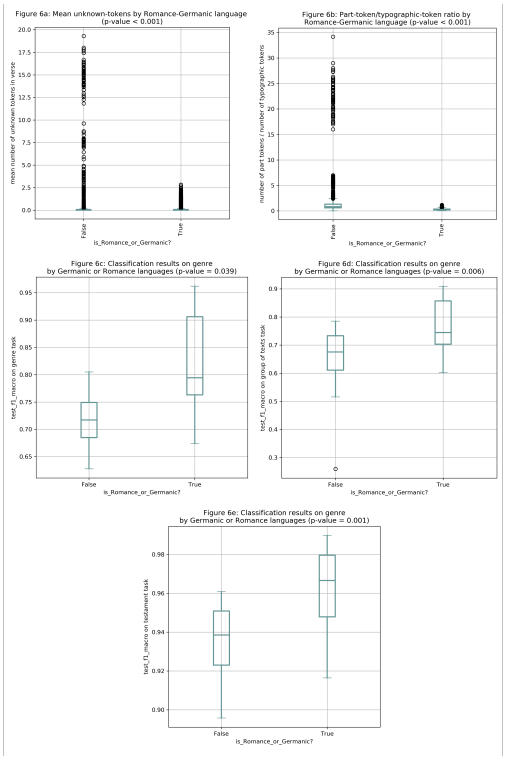
Second, we hypothesize that the Wikipedia sizes correlate with the metrics. While this is rejected for tokenization (Figures 4a-b), the classification results show statistical moderate correlations (Figures 4c-e). Despite having smaller or no Wikipedia, some languages obtain good overall results, such as Haitian, Jamaican, Gothic or Esperanto.



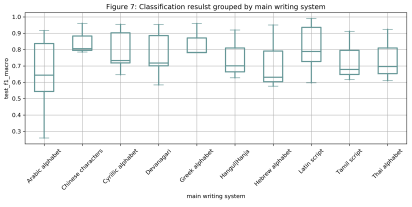
Consequently, in our third hypothesis, we expect texts in Romance and Germanic languages to score better than the rest of the languages. Figure 5 shows the classification results.



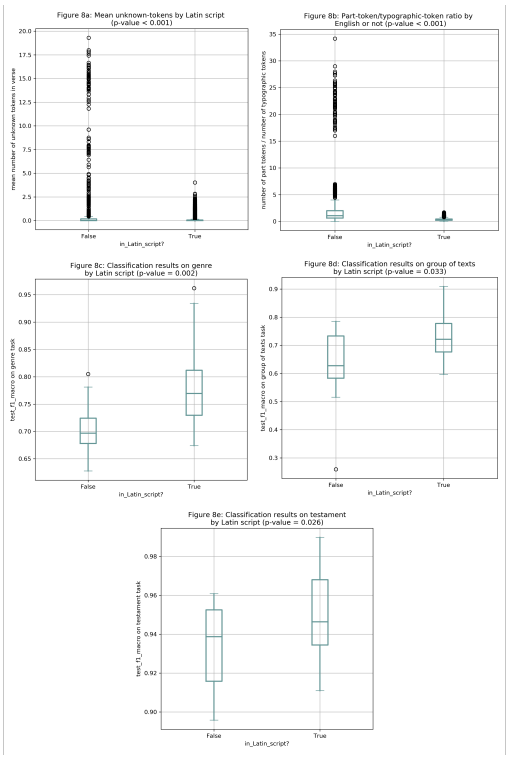
When compared in binary groups, Romance and Germanic languages do obtain better results than the rest for the five metrics (Figures 6a-e).



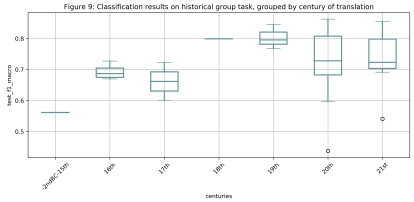
Fourth, based on Muller et al.(2021), we hypothesize that translations using Latin script will score better than translations in other scripts. Figure 7 shows the number of unknown-tokens, notably high for Coptic and Syriac scripts.



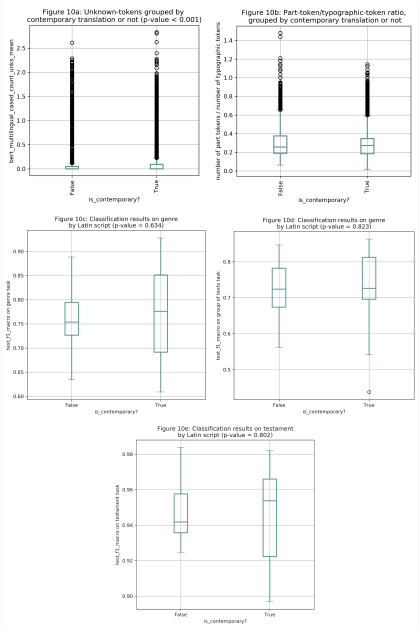
When compared in binary groups, the five criteria support this hypothesis (Figures 8a-e). This might explain the good overall results for some low-resource languages with small or no Wikipedia, but written in Latin script.



Fifth, we expect better results for contemporary translations (20th and 21st century), as that constitutes both the core of the pre-training and Bibles datasets. We obtained the year of publication for 68% of the corpus. Figure 9 shows a positive trend over time, reaching stability after the 18th century.



However, when compared in binary groups (20th-21st vs. rest), this is only supported by the unknown-tokens metric (Figures 10a).



Conclusions

Our evaluation effectively exhibits biases in favor of some high-resource families of languages (Germanic, Romance), although other languages (Haitian, Jamaican, Esperanto) perform reasonably well. At the historical level, the scores are high and stable not only for the 20th and 21st century as hypothesized, but since the 18th century, probably due to a more consistent spelling since then. However, the strongest bias is not towards language or period, but toward (Latin) script. Therefore, transliteration into Latin script and modernization could be mandatory steps for many DH corpora interested in using mBERT. The DH community needs to discuss in which cases modernization and transliteration are acceptable and inwhich ways these limitations could be effectively mitigated.

Bibliography

Alammar, J.(2018). The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning) *Visualizing Machine Learning One Concept at a Time* http://jalammar.github.io/illustrated-bert/ (accessed 7 November 2021).

Bender, E. M., Gebru, T., McMillan-Major, A. and Mitchell, M.(2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Virtual Event Canada: ACM, pp. 610–23 doi:10.1145/3442188.3445922. https://dl.acm.org/doi/10.1145/3442188.3445922 (accessed 3 November 2021).

Burnard, L., Schöch, C. and Odebrecht, C.(2021). In search of comity: TEI for distant reading. *Journal of the Text Encoding Initiative*(Issue 14). Text Encoding Initiative Consortium doi:10.4000/jtei.3500. https://journals.openedition.org/jtei/3500 (accessed 10 September 2021).

Calvo Tello, J.(2020). What is a Genre? A Graph Unified Model of Categories, Texts, and Features. Ottawa: ADHO https://hcommons.org/deposits/item/hc:31713/ (accessed 12 October 2020).

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., et al.(2021). Extracting Training Data from Large Language Models. *ArXiv:2012.07805 [Cs]* http://arxiv.org/abs/2012.07805 (accessed 7 November 2021).

Christodouloupoulos, C. and Steedman, M.(2015). A massively parallel corpus: the Bible in 100 languages. *Language Resources and Evaluation*, **49**(2): 375–95 doi:10.1007/s10579-014-9287-y.

Devlin, J.(2018). *Multilingual BERT Documentation*. https://github.com/google-research/bert/blob/a9ba4b8d7704c1ae18d1b28c56c0430d41407eb1/multilingual.md.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.(2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]* http://arxiv.org/abs/1810.04805 (accessed 22 October 2021).

Ehrmanntraut, A., Hagen, T., Konle, L. and Jannidis, F.(2021). Type- and Token-based Word Embeddings in the Digital Humanities. http://ceur-ws.org/Vol-2989/long_paper35.pdf (accessed 12 October 2021).

Fonteyn, L.(2020). What about Grammar? Using BERT Embeddings to Explore Functional-Semantic Shifts of Semi-Lexical and Grammatical Constructions. https://2021.computational-humanities-research.org/cfp/ (accessed 12 October 2021).

Han, X. and Eisenstein, J.(2019). Unsupervised Domain Adaptation of Contextualized Embeddings for Sequence Labeling. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 4238–48 doi:10.18653/v1/D19-1433. https://aclanthology.org/D19-1433 (accessed 7 November 2021).

Holmes, D. I.(1991). Vocabulary Richness and the Prophetic Voice. *Literary and Linguistic Computing*, **6**(4): 259–68 doi:10.1093/llc/6.4.259.

Jannidis, F. and Konle, L.(2020). Domain and Task Adaptive Pretraining for Language Models. https://dh-abstracts.library.cmu.edu/works/10214 (accessed 12 October 2021).

Lee, J. and Yeung, C. Y.(2016). An Annotated Corpus of Direct Speech. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), pp. 1059–63 http://www.lrec-conf.org/proceedings/lrec2016/pdf/1061_Paper.pdf (accessed 20 April 2019).

Muller, B., Anastasopoulos, A., Sagot, B. and Seddah, D.(2021). When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 448–62 doi:10.18653/v1/2021.naacl-main.38. https://aclanthology.org/2021.naacl-main.38 (accessed 8 November 2021).

Neumann, K. J.(1990). *The Authenticity of the Pauline Epistles in the Light of Stylostatistical Analysis*. (Society of Biblical Literature Dissertation Series 120). Atlanta, GA: Scholars Press.

Novakova, I. and Siepmann, D.(2020). *Phraseology and Style in Subgenres of the Novel: A Synthesis of Corpus and Literary Perspectives*. Cham, Switzerland: Palgrave Macmillan.

Odebrecht, C., Burnard, L., Colorado, B. N. and Schöch, C.(2019). European Literary Text Collection (ELTeC): Release with 10 collections of at least 50 novels. Zenodo doi:10.5281/ZENODO.4274954. https://zenodo.org/record/4274954 (accessed 17 February 2021).

Pagel, J., Sihag, N. and Reiter, N.(2021). Predicting Structural Elements in German Drama. http://ceur-ws.org/Vol-2989/long_paper35.pdf (accessed 12 October 2021).

Radday, Y. T.(1973). *The Unity of Isaiah in the Light of Statistical Linguistics*. (Collection Massorah. Série 2, Etudes Quantitives et Automatisées 1). Hildesheim: Gerstenberg.

Resnik, P., Olsen, M. B. and Diab, M.(1999). The Bible as a Parallel Corpus: Annotating the 'Book of 2000 Tongues'. *Computers and the Humanities*, **33**(1): 129–53 doi:10.1023/A:1001798929185.

de la Rosa, J., Pérez, Á., Sisto, M. de, Hernández, L., Díaz, A., Ros, S. and González-Blanco, E.(2021). Transformers analyzing poetry: multilingual metrical pattern prediction with transfomer-based language models. *Neural Computing and Applications*doi:10.1007/s00521-021-06692-2. https://doi.org/10.1007/s00521-021-06692-2 (accessed 22 November 2021).

Schöch, C., Erjavec, T., Patras, R. and Santos, D.(preprint). Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives. *Modern Languages Open*doi:10.5281/zenodo.4742420. https://zenodo.org/record/4742420 (accessed 10 September 2021).

Sims, M., Park, J. H. and Bamman, D.(2019). Literary Event Detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 3623–34 doi:10.18653/v1/P19-1353. https://aclanthology.org/P19-1353 (accessed 7 November 2021).

Tunstall, L., Werra, L. von and Wolf, T.(preprint). *Natural Language Processing with Transformers: Building Language Applications with Hugging Face.*Sebastopol: O'Reilly Media.

Underwood, T.(2021). Mapping the Latent Spaces of Culture *Using Large Digital Libraries to Advance Literary History*. Humanities Commons https://tedunderwood.com/2021/10/21/latent-spaces-of-culture/ (accessed 3 November 2021).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I.(2017). Attention Is All You Need. *ArXiv:1706.03762 [Cs]* http://arxiv.org/abs/1706.03762 (accessed 8 November 2021).

Walczak, B.(2015). Role of the Bible in the development of languages and linguistics. An outline of the issue. *Forum Lingwistyczne*(2). Wydawnictwo Uniwersytetu Śląskiego / University of Silesia Press https://www.journals.us.edu.pl/index.php/FL/article/view/6066 (accessed 15 November 2021).

Zimmermann, R.(2010). Theologische Gattungsforschung. In Zymner, R. (ed), *Handbuch Gattungstheorie*. Stuttgart: Verlag J.B. Metzler, pp. 302–05.

Archive of the Digital Present (ADP), COVID-19 Period: Collecting and Visualizing Metadata of Online Literary Events Hosted in Canada, March 2020 - September 2021

Camlot, Jason
jason.camlot@concordia.ca
Concordia University, Canada

Neugebauer, Tomasz
Tomasz.Neugebauer@concordia.ca
Concordia University, Canada

Berrizbeitia, Francisco
francisco.berrizbeitia@concordia.ca
Concordia University, Canada

Joseph, Ben
rohan_ben_joseph@sfu.ca
Simon Fraser University, Canada

Bustamante, Alexandre
alexandre.bustamante@gmail.com
Concordia University, Canada

Gandham, Sukesh
sukesh.gandham@gmail.com
Concordia University, Canada

Archive of the Digital Present for Online Literary Performance in Canada (COVID-19 Pandemic Period) is a research and development project that arises out of the need to address foundational, practical and theoretical research questions about the impact of the recent (and ongoing) COVID-19 pandemic, and attendant social disruptions and restrictions, upon literary communities in Canada through the study of organised literary events as they have occurred online since March 2020.

The papers that constitute this panel focus on the design and development work pursued in building a searchable, open access database and directory – The Archive of the Digital Present (ADP) – to allow scholars, literary practitioners, and the public to gain knowledge about the nature and significance of literary events (online, hybrid, and in-person) that have occurred during the pandemic period, through the collection and structuring of metadata, and, in some cases, with direction to audiovisual (AV) documentation of the events themselves as they were held using platforms such as Zoom and YouTube.

Our papers explain key facets of development by presenting approaches to (1) data collection and structuring, (2) stack development, (3) data visualisation, and (4) front end design, that have emerged through the process of community and user-oriented design research and development used to create the ADP.

Finding and Structuring Metadata about Pandemic Literary Events (Jason Camlot)

Public readings represent a significant form of literary communication, dissemination, circulation and community-building. The study of literary performances, events and activities through audiovisual media documentation, digital images (posters) and textual records, raises important new questions about literary work as it acts *in situ* among artists and audiences (Camlot, Fong and Shearer). Such materials and the events they document reveal unique traces of sociality and affective response in literary exchange, foreground tonal and performative aspects of cultural transmission, document formations of literary community in action, and highlight the mediated nature of such events as they first occurred and as we subsequently access them through archived recordings. Focusing on the presentation of digital documentation and records of online events of the COVID-19 pandemic period, the ADP is designed to help us understand the impact of pandemic disruptions on literary communities in Canada.

The ADP project necessarily began with questions about the data we were seeking to collect. In February 2021 we performed a preliminary analysis of online and social media postings for listings of literary events hosted in Canada. This revealed 77 discrete organisers of over one thousand (1,011, to be exact) literary events between 20 March 2020 - 31 December 2020. This list served as the starting point for an expanded catalogue of events, and for team discussions about the nature and number of metadata fields we would use. We proceeded by adapting extant categories of the SpokenWeb metadata schema that has been designed for the description of historical literary audio recordings. This allowed us to repurpose the backend of the Swallow Metadata Ingest System (*Swallow*), built for metadata management of historical research collections, through the development of a crosswalk that best serves the goals of data collection for ADP. Storing the metadata as unstructured JSON, *Swallow* makes it possible to quickly generate and modify a cataloguing interface, and changes in metadata schemas, and this flexibility has proven useful for the iterative design and feedback process we have pursued in ADP UX development. Data fields we have shaped for this project include categories related to Title, Creator/Contributor, Language, Production Context, Genre, Duration, Date, Location, Online Platform, and Contents, among others.

This section of our presentation presents our ongoing methods of discovering events to be included in the ADP database, explains the rationale of our selection of metadata categories and our approach to structuring those fields, and raises some of the philosophical and ontological questions that have arisen in the process of abstracting the complex and mediated literary activities of the pandemic period into categories of searchable data. Our methods of data collection extended beyond the analysis of social media notices of events to the identification and research of key organisations involved in hosting events, and to crowdsourcing calls within diverse literary communities. Our selection of the specific fields identified were based on feedback from researchers and practitioners about the kinds of information they would seek from a database such as this, and on our goals in data visualisation, discussed below.

Philosophical and ontological questions raised by our data determination and collection process included questions about the nature of an event, the roles of participants in relation to categories such as creator and contributor, and generic categories that include or preclude the overarching category of the "literary" event, and the status of such events as entities within a data structure.

Stack Development and the Rationale of a Headless CMS (Ben Joseph, Francisco Berrizbeitia)

In this section we present the software stack chosen to develop the solution and the rationale behind it. The right technical stack is, to a great extent, the key to a project's success, while the wrong choice of web application development technologies may be a reason for failure. Our stack had to manage the trade-offs for processing heavy loads and maintaining a low latency with high responsiveness.

The data was collected using *Swallow* (Camlot et al. 2020), an open-source metadata management system developed by the SpokenWeb partnership. Interacting with this backend system, cataloguers collected and compiled the information of the different events. This corpus of data is then exported as whole and ingested by the Strapi headless content management system (https://strapi.io/) as shown in (**Figure 1**. MeiliSearch (https://www.meilisearch.com) provides advanced search functionalities and Strapi allows for digital asset management while providing a backend to the presentation layer. The presentation layer was developed using Angular (https://angular.io/). The component-based structure of Angular makes specific sections of code highly reusable across our application and for future frontend developments. The AngularJS layer interacts with the data sources via a GraphQL (https://graphql.org/) thus ensuring data consistency.

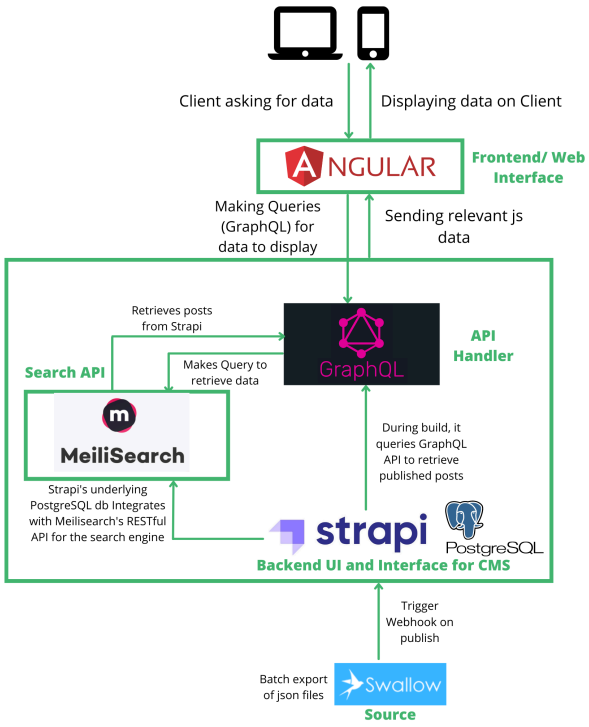


Figure 1.
Software stack schema and information flow.

Visualizing Time, Place and Relations of Literary Activity in a Pandemic (Tomasz Neugebauer, Sukesh Gandham)

A user survey identified the host location of the event as one of the more important elements of event description, along with event type, contributors, links to recordings, and the titles of texts read. The most important metadata elements for navigation of content were identified as: names of participants, names of organizers and event, titles of texts read/performed/discussed, and date/time of event. This lead to an initial front page design for the ADP site that included three visualizations of the data: A) Timeline: a spiral histogram (Condegram) showing the number of events that took place on specific dates during the pandemic period B) Connections: a network graph showing adjacency relations between contributors and events, using hierarchical edge bundling based on the events' organizations, and C) Places: geographical map showing number of events that took place in each of the events' host cities. These visualizations touched on all the important metadata elements identified in the user surveys, except for the titles of texts read/performed.

During the user-centered design process, a usability session evaluating some proposed wireframes for the

site with scholars from the Spokenweb network, event organizers, practitioners and students confirmed that there is interest and enthusiasm for the data visualizations on the interface, especially for any interactive functionalities of these. To keep the visualizations interactive, we decided to treat them as tools that navigate into the data. For example, clicking on a name of a contributor or event in the edge bundling visualization leads the user to the browse view of the search results in the dataset for that name. Similarly, clicking on a bar for a specific date on the condegram, or the city name on the map visualization, leads the user to the browse view of events that took place on that date or that were hosted in that city.

For the implementation of these visualizations, keeping complexity in check, we wanted the same code base for all three of them. We also wanted a solution that is based on open-source code that is free from license restrictions on usage and sharing, which eliminated solutions such as amCharts (https://www.amcharts.com/). We chose the D3.js (https://d3js.org/) JavaScript library as our code base to generate Scalable Vector Graphics (SVG) that we present with custom HTML5 and Cascading Style Sheets (CSS). We added the Bootstrap library (https://getbootstrap.com/docs/3.4/) for responsive functionality. The D3 Gallery (https://observablehq.com/@d3/gallery) served as an excellent starting place for choosing the closest existing examples to build on. We used Python to transform the ADP JSON data from our Swallow Metadata Management System into data shapes that are required for the three visualizations using D3.

User-Oriented Design and Aesthetics for a Pandemic-Period Website (Alexandre Bustamante)

The avenue for building the front-end of ADP was, from an early phase, directed as a user-centred project. After the design was initiated, discussions evolved for experimenting with the lines of a Participatory Design (PD) approach. A user-centred methodology is considered executed "on behalf of users" while PD approaches design "with the users" (Spinuzzi 2005, 165), We chose this option to have ensure greater involvement of the stakeholders of the ADP in all development phases of the front-end design. We organized a series of workshops, PD activities and user-experience research surveys distributed to invited participants and the direct team of the front-end development, with the project designer playing a facilitating rather than individual authorial role. The workshops adopted established design methods (Martin 2012) such as personas and user journeys for delineating stakeholder needs. Card sorting, tree-testing and first-click tests were used to create and confirm an informational architecture (figure 2). We adapted to the added challenge of conducting workshops remotely due to the pandemic. This led to the exploration of available tools, such as *Miro Boards* (https://miro.com), (*Google Jamboards* ((https://jamboard.google.com/) and *Optimal Workshop* (https://www.optimalworkshop.com).

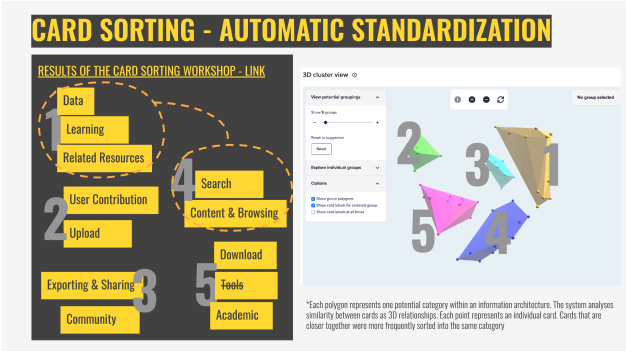


Figure 2
Results from an online workshop, with automated interpretation of a card sorting exercise compiled by the user-experience tool Optimal Workshop. *Screenshot by the authors.*

The initial outcome of the PD process was the creation of a three-level information architecture for the prototype of the front-end design: starting at the first level with an overall glimpse at the directory of events on the home page which presents content through visualizations and browsing functions. From the homepage, the user moves to a second level where content is presented on either a dashboard or in the form of lists in pre-established categories, displaying more information and details and introducing different filters for the content. Finally, the third and final level of the information architecture is detailed access to the metadata, which can be visualized, exported or shared. The front-end design allows a search function to be performed at all three levels, for direct access to the directory of events.

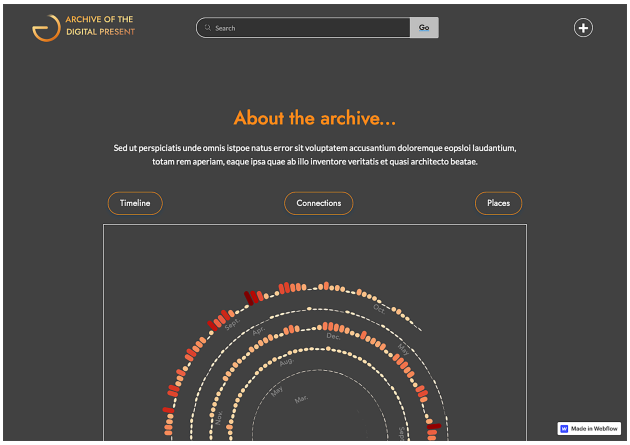


Figure 3
Latest front-end design for the ADP website. A glimpse of the landing page of the Archive. Screenshot by the authors.

Once the information architecture was established, the participatory approach to design was expanded to inform the final look of the prototype (figure 3). Team members and stakeholders were invited to guide the visual design by reflecting on their perception of the pandemic and providing keywords that capture the experience of living through this period. The final prototype was then designed to reflect on these shared perceptions, drawing from participant's contributions to inspire a mood board for the design work, which directed the design decisions to achieve a final result that aims to be characteristic of its particular time.

Bibliography

Angular: The modern web developer's platform! (n.d.). https://angular.io/ (accessed 27 July 2022).

Camlot, Jason. (2013). The Sound of Canadian Modernisms: The Sir George Williams University Poetry Series, 1966-1974. *Journal of Canadian Studies / Revue d'études canadiennes,* 46(3): 28-59.

Camlot, Jason , Neugebauer, Tomasz and Berrizbeitia, Francisco. (2020). Dynamic Systems for Humanities Audio Collections : The Theory and Rationale of Swallow. *DH2020 (Digital Humanities 2020 Virtual Conference), 23 July 2020, Ottawa, Canada.* https://spectrum.library.concordia.ca/id/eprint/987014/ (accessed 10 April 2022).

Fong, Deanna, and Karis Shearer. (2018). Gender, Affective Labour, and Community Building Through Literary Audio Artifacts. *SpokenWebBlog*. https://spokenweb.ca/spokenweblog/ (accessed 20 April 2022).

GraphQL: A query language for your API. (n.d.). GraphQL. https://graphql.org/ (accessed 9 December 2021).

Martin, Bella, and Bruce M Hanington. (2012). *Universal Methods of Design: 100 Ways to Research Complex Problems, Develop Innovative Ideas, and Design Effective Solutions.* Beverly, MA: Rockport.

Meilisearch (n.d.). Meilisearch Github Repository. https://github.com/meilisearch/meilisearch (accessed 27 July 2022).

Spinuzzi, Clay. (2005). The Methodology of Participatory Design. *Technical Communication*, 52(2): 163–74.

Strapi: Open source Node.js headless CMS to easily build customisable APIs. (n.d.). Strapi Github Repository. https://github.com/strapi/strapi (accessed 9 December 2021).

Textual, Metrical and Musical Stylometry of the Trouvères Songs

Camps, Jean-Baptiste
jean-baptiste.camps@chartes.psl.eu
École nationale des chartes | Université PSL, France

Chaillou, Christelle
chaillouchristelle@gmail.com
CESCM, CNRS, Poitiers, France

Mariotti, Viola
viola.mariotti.maritem@gmail.com
CESCM, CNRS, Poitiers, France

Saviotti, Federico
federico.saviotti@unipv.it
Università degli Studi di Pavia

The lyrical tradition and the *Chansonnier du roi*

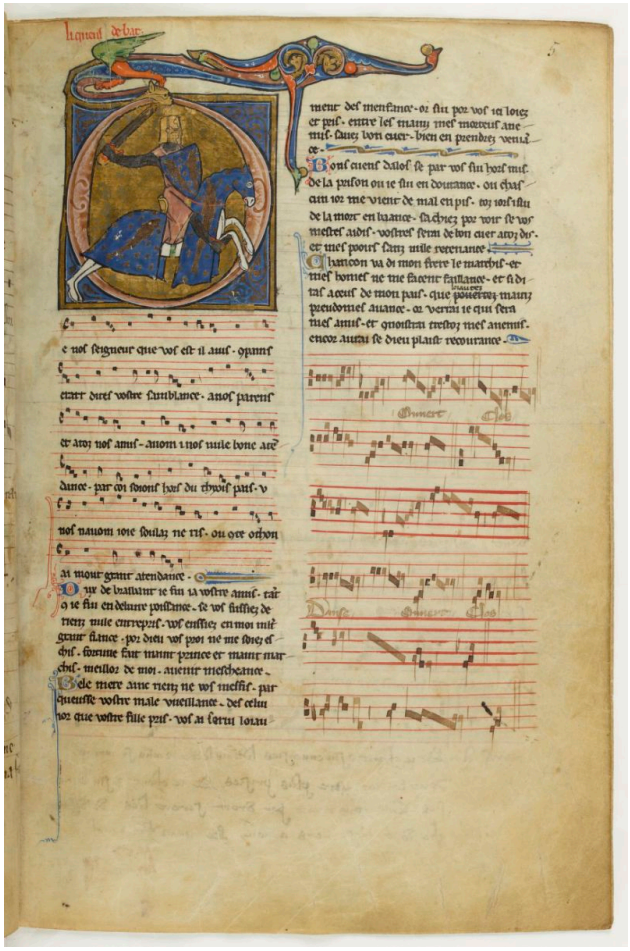


Figure 1:
BnF, fr. 844, fol. 29 (manuscrit du roi), beginning of the section for the texts of the count of Bar.

In the landscape of medieval French texts' authorship, the lyrical poetry of the *trouveres* and *trouveresses* has the aspect of an exception: contrarily to what can be observed elsewhere, anonymity is rare, and the manuscripts mostly take care to give a clear (though often conflicting between them) attribution, even up to the point of adopting a material organisation by author (fig. 1). Yet, very little is known about the attribution of the melodies that accompany many of these texts: how often are they original? Were they composed by the *trouvere* himself, or by one or several other composers? Is there even a unity behind the different melodies of the work of a given author?

To answer these questions, we propose to experiment in the cross examination of the text, the metrical composition and the melodies of a corpus of lyrical texts with the methods of computational stylometry. We will focus on the *Manuscrit du Roi* (Paris, BnF fr. 844), one of the earliest (1260-1270) and richest sources for Romance lyric poetry with its 602 compositions. The literary, musical and linguistic diversity of its contents (profane songs of French trouvères and Occitan troubadours, French motets, instrumental works and Latin sacred compositions) makes it a unique case-study.

The contents of this manuscript were acquired thanks to a workflow going from layout analysis, handwritten text recognition to digital editing and ultimately stylometric analysis. Text and music are encoded following the *Guidelines* of, respectively, the *Text Encoding Initiative* (TEI, 2020) and the *Music Encoding Initiative* (MEI, 2020), resulting in a textual and musical edition (See, for a detailed presentation of the ongoing editorial process, Camps *et al.*, 2021a).

For the needs of the textual analysis of the text, the HTR output, after careful human correction, was then segmented into words, normalised (abbreviations expanded, allographs normalised), lemmatised and POS-tagged using deep learning methods (Table 1). Even if the scores obtained are high, cumulative errors through the different steps might remain an issue.

Table 1: Word accuracy is indicated with respect to the in-domain test set (samples set apart from training material).

Step	Word acc	Engine	Data/Model
segmentation	96.71	Boudams (Clérice, 2020)	10.5281/zenodo.6500604
normalisation	98.01	Pie (Manjavacas *et al.*, 2019)	10.5281/zenodo.6500649
lemmatisation	97.66	Pie (*ibid.*)	10.5281/zenodo.4320487
pos-tagging	97.55	Pie (ibid.)	10.5281/zenodo.4320487

The availability of a complete transcription of both text and musical notations paves the way for the stylometric analysis of the songs of the *trouvères* and *trouveresses*, at a level impossible until now. This is a critical issue, because disputed attributions are very numerous inside the Old French Lyrical tradition (Gatti, 2019), yet it poses specific challenges, because the components are very short (often less than 50 verses), idiolect of the text appears rather homogeneous – sharing as they are on surface the same elitist cultural tradition –, while, in the meantime, the manuscript tradition creates noise (linguistic and substantial variants, even between scribes of the same manuscript). Moreover, the field of musical stylometry is still quite new,

and attribution of the text and of the melody have rarely - if ever - been addressed as a whole in this tradition.

By default, scholars generally postulate that the melody and the text of a song were composed by the same person. Yet the attested musical elements (for exemple : musical curbe, modality, intervals, ambitus, motives, melismas and musical form) have usually been considered as much more influenced by the tradition than the textual ones. Nonetheless, it will be worth testing the hypothesis that song attribution may be consistently built on both textual and musical elements.

Specifically, the Chansonnier du roi offers an interesting dossier, because it contains a subunit dedicated to the collected work of a single author (Thibaut de Champagne so-called *Liederbuch*). Thibaut is perhaps the most prominent of all trouvères (Barbieri, 1999), and the attribution of several songs to him is still disputed (Wallensköld, 1925; Callahan, 2010).

Stylometric analysis of the text

In order to give new insights into these disputed attributions, we performed several stylometric analyses of the text and the music.

Stylometric analysis of the text has initially been done using features robust to noise and short text length, in particular character 3-grams (Camps *et al.*, 2021a). Both exploratory and supervised analyses, the latter using a linear SVM , were performed to shed more light on the attribution of these components. To compensate for the imbalance between the available samples for all four authors, various upsampling and downsampling strategies (alone or in combination) were experimented, and Tomek links removal (Tomek, 1976), in combination with class weights, was found to be the best performing. Evaluation was done using a leave-one-out approach, and the models reached a global accuracy of 0.79, with a F1 score of 0.93 for Thibaut and 0.76 for Gace (Table 2).

Table 2: Results of the leave-one-out evaluation on the text of the songs, using character 3-grams, lemmas, part-of-speech 3-grams and lemmatised function words

	character 3-grams			
	precision	**recall**	**f1**	**support**
Blond	0.85	0.52	**0.65**	21
Gace	0.69	0.86	**0.76**	43
Gaut	0.71	0.60	0.65	20
Thib	0.93	0.93	**0.93**	45
—				
accuracy			**0.79**	129
macro avg	0.79	0.73	**0.75**	129
weighted avg	0.80	0.79	**0.79**	129

	Lemmas (all)			
	precision	**recall**	**f1**	**support**
Blond	0.77	0.48	0.59	21
Gace	0.70	0.65	0.67	43
Gaut	0.85	0.55	**0.67**	20
Thib	0.65	0.91	0.76	45
—				
accuracy			0.70	129
macro avg	0.74	0.65	0.67	129
weighted avg	0.72	0.70	0.69	129

	POS 3-grams			
	precision	**recall**	**f1**	**support**
Blond	1.00	0.24	0.38	21
Gace	0.43	0.49	0.46	43
Gaut	0.17	0.05	0.08	20
Thib	0.43	0.67	0.53	45
—				
accuracy			0.44	129
macro avg	0.51	0.36	0.36	129
weighted avg	0.48	0.44	0.41	129

	Lemmas (function words)			
	precision	**recall**	**f1**	**support**
Blond	0.59	0.48	0.53	21
Gace	0.45	0.40	0.42	43
Gaut	0.25	0.30	0.27	20
Thib	0.62	0.69	0.65	45
—				
accuracy			0.50	129
macro avg	0.48	0.47	0.47	129
weighted avg	0.50	0.50	0.50	129

Yet, the extraction of the features with the highest coefficients in the SVM models show a more contrasted result. For an author like Gace Brulé, the 3-grams seem to reveal aspects significant for our stylometric analysis and reflecting lexical and morphological preferences of the author. For instance, the high-frequency of *apl*, *usp* depends mainly of the syntagm *a plaisir* (written as one single word) and the words *souspir*/ *souspirer* respectively, which seem

to be typical of Gace's idiolect, whereas the low-frequency *ete* is consistent with the raffinate poet's avoidance of the diminutive nouns and adjectives in - *ete*, as marks of a popular register, and his apparent rare use of some abstract nouns like *faussete*. Yet for Thibaut's corpus, the analysis of the most and least recurrent 3-grams seems to include features that are not of an authorial nature but reflect the particular graphemic habits of the copyist of the part of the manuscript known as Thibaut's *Liederbuch* as opposed to the scribes at work in the rest of the manuscript: use of ‹o› instead of ‹ou› in particular in words like *po(u)r*, *to(u)z*, *amo(u)rs*, use of ‹u›, not ‹v› at the beginning of words (this last feature being even of a more graphetic nature).

To bypass this inconvenient, analyses were repeated, using lemmas, part-of-speech 3-grams and lemmatised function words. Results actually were found to be less accurate (Table 2), which can either point to a less informative nature of these features (i.e., excessive suppression of information and noise generation through lemmatisation, tagging and function words selection), or to the fact that some of the performance of the 3-grams model is due to correlation between scribal practice and authorial units. If that were true, normalisations could both potentially decrease the performance of the model and increase the authorial nature of the features it uses. The relative better performance of lemmas (function and content words) in comparison to part-of-speech or function lemmas could also point towards the importance of lexical choices to discriminate between authors in the training material.

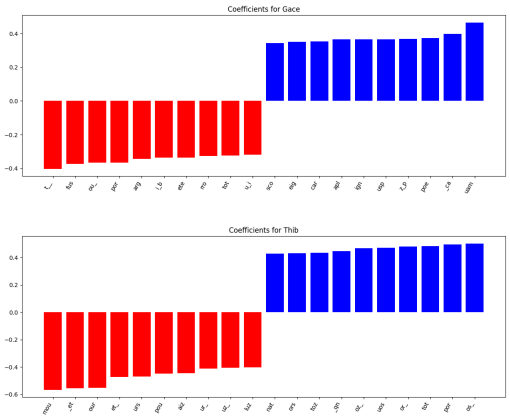


Figure 2:
Features with the 10 highest positive and negative coefficients in the linear SVM models for Gace Brulé and Thibaut de Champagne; the higher the coefficient, the more the feature would contribute to a placement on one side of the separating hyperplane or the other

Stylometric analysis of the music

Musical stylometry (especially of the *trouvères*) is a much less waymarked field. It is supposed that all composers integrate some little elements consciously or unconsciously in their compositions. A recent study observed that the troubadour Bernard de Ventadorn used in most of this songs three ascendants sounds at the beginning of verses (Chaillou-Amadieu 2016). Such methodology can be automated and show promise for the joint stylometry of text and music. In order to identify relevant features, experiments have been conducted with a variety of variables: notes, intervals, octaves n-grams, with the inclusion of syllable boundaries in the absence of noted measure, all extracted from MEI neumes notation. They were done on two different dataset: a small, controlled dataset (3x10 melodies of supposedly three authors), and an enlarged and more difficult dataset (117 melodies, with imbalanced training set of six authors).

Table 3: Benchmark of musical features for musical stylometry

	Smaller set					
n	*1*	*2*	*3*	*4*	*5*	*6*
notes	0.67	**0.80**	0.73	0.67	0.63	0.70
octaves	0.57	0.53	0.57	**0.60**	0.57	**0.63**
intervals	0.77	**0.80**	0.73	0.70	0.70	0.53

	Larger set					
n	*1*	*2*	*3*	*4*	*5*	*6*
notes	0.29	**0.38**	**0.38**	**0.38**	0.26	0.26
octaves	0.31	0.44	**0.47**	0.41	0.38	**0.47**
intervals	0.37	0.40	0.41	0.43	**0.46**	0.40

Intervals or notes bigrams have proved to be the most efficient on the smaller set, while results on the larger set remain more difficult to interpret. A visualisation on the smaller set shows how these features can structure a vector space of partly authorial data clouds (fig.), yet a benchmark of the various types of features for supervised learning gives results whose interpretation is more elusive (Table 3), apart from the obvious impact of availability of training samples (Table 3).



Figure 3:
First factor plane from principal component analysis of notes and intervals bigrams (smaller set)

Table 4: Detailed leave-one-out evaluation of the SVM linear models on intervals bigrams from the larger set (with random downsampling)

	precision	recall	f1	support
BlonNes	0.21	0.19	0.20	16
GaBru	0.41	0.35	0.38	34
GautDarg	0.29	0.31	0.30	16
GuilVinier	0.29	0.33	0.31	12
Moniot	0.50	0.75	0.60	12
ThibChamp	0.68	0.63	0.65	27
—				
accuracy			0.43	117
macro avg	0.40	0.43	0.41	117
weighted avg	0.43	0.43	0.42	117

Discussion and further research

Initial results show good and average performance on the attribution of texts and melodies, with an interesting better performance for Thibaut and Gace, our two main target authors.

Yet, important work remains to be done, on one hand, to measure the relevancy of different features both for textual and musical stylometry, going beyond simple metrics. Indeed, our work results stress the necessity to go beyond bare metrics in the evaluation of supervised model training, to actually consider the precise features on which the model bases its attributions.

Moreover, we aim to explore features that are at the intersection of text and music, such as recurring associations between certain musical and textual features, as well as metric patterns.

Bibliography

Callahan, C. (2010). 'Thibaut de Champagne and Disputed Attributions: The Case of MSS Bern, Burgerbibliothek 389 (C) and Paris, BnF fr. 1591(R)'. *Textual Cultures*, 5(1). Indiana University Press: 111–32 doi: 10.2979/tex.2010.5.1.111.

Camps, J. B., Clérice, T., & Pinche, A. (2021a). 'Noisy medieval data, from digitized manuscript to stylometric analysis: Evaluating Paul Meyer's hagiographic hypothesis'. *Digital Scholarship in the Humanities*, 36(Supplement_2), ii49-ii71.

Camps, J.-B., Chaillou, C., Mariotti, V. and Saviotti, F. (2021b). 'Editing and Attributing Musical Texts: the Chansonnier du Roi and the MARITEM Project'. *EADH2021: Interdisciplinary Perspectives on Data, 2nd International Conference of the European Association for Digital Humanities, Krasnoyarsk*, 2021 https://halshs.archives-ouvertes.fr/halshs-03260116/document.

Chaillou-Amadieu, C. (2017). *Philologie et musicologie. Les variantes musicales dans les chansons de troubadours. In Les noces de Philologie et de musicologie. Textes et musiques du Moyen Âge*, ed. C. Cazaux-Kowalski and al. Paris: Classiques Garnier, 2017, p. 69-95.

Clérice, T. (2020). 'Evaluating Deep Learning Methods for Word Segmentation of Scripta Continua Texts in Old French and Latin'. *Journal of Data Mining & Digital Humanities*, 2020.

Clérice, T., Camps, J.-B. and Pinche, A. (2019). Deucalion, Modèle Ancien Francais (0.2.0). Zenodo doi: 10.5281/zenodo.3237455..

Dees, A. (1987). *Atlas des formes linguistiques des textes littéraires de l'ancien français*, Tübingen, Niemeyer.

Gatti, L. (2019). *Repertorio delle attribuzioni discordanti nella lirica trovierica*, Roma, Sapienza Università Editrice (Online: http://www.editricesapienza.it/sites/default/files/5850_Gatti_Lirica_trovierica_interior_OA.pdf).

Manjavacas, E., Kádár, Á. and Kestemont, M. (2019). 'Improving Lemmatization of Non-Standard Languages with Joint Learning'. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies Minneapolis, p. 1493–503, doi: 10.18653/v1/N19-1153.

Music Encoding Initiative (2020). *Guidelines*. Mainz https://music-encoding.org/ (accessed 8 January 2021).

TEI Consortium (2020). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. https://tei-c.org/Guidelines/ (accessed 9 May 2015).

Tomek, Ivan (1976). 'Two modifications of CNN', *IEEE Trans. Systems, Man and Cybernetics*, 6:769–772.

Troubadour Melodies Database, http://www.troubadourmelodies.org/melodies.

Wallensköld, A. (1925), ed.. Thibaud IV (1201-1253 ; comte de Champagne): Les chansons de Thibaut de Champagne, roi de *Navarre*. https://gallica.bnf.fr/ark:/12148/bpt6k53236 (accessed 8 January 2021).

Data Diversity in handwritten text recognition: challenge or opportunity?

Camps, Jean-Baptiste
jean-baptiste.camps@chartes.psl.eu
École nationale des chartes | Université PSL, France

Pinche, Ariane
ariane.pinche@chartes.psl.eu
École nationale des chartes | Université PSL, France

Stutzmann, Dominique
dominique.stutzmann@irht.cnrs.fr
Institut de recherche et d'histoire des textes/CNRS

Vernet, Marguerite
marguerite.vernet@chartes.psl.eu
École nationale des chartes | Université PSL, France

Vidal-Gorène, Chahan
chahan.vidal-gorene@chartes.psl.eu
École nationale des chartes | Université PSL, France

Introduction

In this paper, we wish to show approaches in handling diversity in larger collections of training data for text acquisition pipelines, specifically handwritten text recognition for medieval manuscripts in Latin and French. Present throughout medieval Europe, Latin is one, if not the most used written language of the time on this continent, while French has known from a relatively early date (around the 12th century judging from preserved manuscripts) a vernacular production that soon became one of the most prominent of Western Europe, influencing the written culture of its neighbours from its central position. Combined, they provide a case study whose diversity and general scope could, we hope, allow to provide results with broader applicability, even beyond medieval Western manuscripts.

Heterogeneity or diversity in the collections can result from intrinsic features (e.g. linguistic, palaeographic, diachronic variation in the sources), but also from extrinsic features (aim and provenance of transcriptions, idiosyncrasies of transcribers…). We propose to approach both types of diversity by reusing several open data sets from various research projects in diverse fields and involving many collaborators. We add a double focus, linguistic (Latin vs. French manuscripts) and graphic (abbreviated vs. normalised transcriptions). We hope to be able to overcome, to some extent, the issue of linguistic diversity and propose a common, modular pipeline for different languages, related but different in their inner structure and declension mechanisms.

When, on the one hand, recent studies focus on "hyperdiplomatic" digital editions to study the production of specific items, the implementation of natural language processing and text mining is commonly based on a normalised text. Instead of aiming at defining a single universal translinguistic transcriptional standard to merge all existing standards – an utopic endeavour, and perhaps even not desirable –, and instead of designing a unified pipeline supported by dedicated libraries (e.g. image > hyperdiplomatic > normalised > lemmatised+POS-tagged > critical text) to constrain all existing editions, we applied a

more modular approach to reuse and pool datasets to train multiple models and design paths more fitted to the variety of goals encountered in medieval studies.

In this attempt, we will strive to answer more specifically the following questions:

1. To what extent can we (and should we) mutualise HTR training material between preexisting datasets and even related languages? (and is it worth the effort?)
2. Are approaches that decompose image to text prediction and further linguistic normalisations (abbreviation expansion for instance) better performing for that goal than straightforward "image to normalised text" approaches?

Diversity in our corpus

Extrinsic diversity: variation in data production

The most obvious source of diversity is artificial, in the sense that it is the result of the production of the data (and particularly of transcription choices) and not of the sources itself.

For this research three macro-datasets, themselves mostly aggregates of smaller micro-datasets, have been used, one French and two Latin.

The French dataset is Cremma-medieval, composed of 17431 lines from eleven Old French manuscripts written between the 13 th and 14 th centuries (Table 1). It is made from pre-existing transcriptions, and sample size is very different from one source manuscript to the other. A graphemic [1] transcription method has been chosen to maintain a many to one mapping between signs in the source and the transcription (abbreviations and their expansions are both kept, *u/v* or *i/j* are not dissimilated), but allographs are normalised (*e.g.*, round and long *s* are both transcribed *s*). Finally, spaces are not homogeneously represented in the ground truth text annotation, with transcribers reproducing the manuscript spacing while others are using lexical spaces. It must be stressed that spaces are the most important source of errors in medieval HTR models (see for instance the model Bicerin, where spaces represent 33.9% of errors Pinche 2021). In this cremma-medieval macro-dataset, several transcriptions from different transcribers, coming from different projects, have been collected.

This diversity is also very present in the Oriflamms macro-dataset, containing 120 111 lines from no less than 779 manuscripts (Table 1). This dataset has been composed along several different projects over a substantial interval of time, and is a mix of aligned preexisting normalised editions (without abbreviations) and graphemic transcriptions (including abbreviations and their expansion). It is composed of both French, Latin and bilingual texts.

Table 1: Composition of the cremma-medieval, Oriflamms and st-victor macro-datasets [For this abstract, only corpora in bold have been used]

Corpus	Editors	Manuscripts	Pages	Lines
Otinel	Camps	2	75	**13568**
Wauchier	Pinche	1	49	**6148**
Maritem	Mariotti	1	18	1026
CremmaLab	Pinche et al.	7	55	13568
—				
Total		11	149	17431
Reg.chancell. Poitou	Guérin	200	1217	30015
Reg.chancell. Paris	Viard	2	29	474
Morchesne	Guyotjeannin et al.	1	189	10394
Cartulaire de Nesle	Hélary	1	117	3899
Chartes Fontenay	Stutzmann	104	104	**1384**
Psautiers	Oriflamms	27	48	5793
PsautierIMS	Stutzmann	48	132	**3145**
MSS dat. lat.	Oriflamms / ecmen	101	101	**2299**
Queste del saint Graal	Marchello-Nizia, Lavrentiev	1	130	10725
BnF fonds fr.	ecmen	159	189	13510
Mss dat. fr.	ecmen	45	55	3355
Album XIIIe.	Careri, et al. + ecmen	52	52	1992
Légende dorée	irht+ ecmen	18	679	31742
Pèlerinage	opvs+ ecmen	20	56	1384
—				
Total		779	3098	120111
Saint-Victor	Vernet	2	54	**12596**

The last macro-dataset Saint-Victor is the most homogeneous, composed of transcriptions from two

Victorine mss, i.e., BnF latin 14588 and BnF latin 14525 written by no less than twelve scribes at the end of the 12 th century and the first part of the 13 th century (Table 1). Both mss have the same type of writing. It has been created during a master's thesis. It is divided into two sub-corpus. A first corpus is transcribed without abbreviations. The transcription uses lexical spaces. It is the most important of the two sub-corpus with 10736 lines. The second sub-corpus consists of a small part of the first (1860 lines), which has been transcribed with abbreviations.

Early tests have shown the tremendous variations in the choice of signs used to transcribed medieval graphemes, in particular abbreviations, including MUFI and out of MUFI characters. For example, the common abbreviative marker has been transcribed alternatively as U+0303 COMBINING TILDE, U+0304 COMBINING MACRON, U+0305 COMBINING OVERLINE, F00A COMBINING HIGH MACRON WITH FIXED HEIGHT (PART-WIDTH), and even, in composition, U+1EBD LATIN SMALL LETTER E WITH TILDE, U+0113 LATIN SMALL LETTER E WITH MACRON, etc. Even when using MUFI (Medieval Unicode Font Initiative), different types of Tironian *et* or *p* flourish can be used. To facilitate machine learning, a conversion table was used to apply a first level of normalisation, and to reduce the 262 preexisting character class to around 30 (Clérice and Pinche 2021).

Intrinsic diversity: variation in language, script and scribal practice

Diversity is also due to linguistic differences inside the corpus, with a main distinction between Latin and French texts, the latter in a variety of regional *scriptae*, including Anglo-French, Eastern (Lorrain) and Picard, and also diachronic variation, from 12 to 14th century.

The variety is also in the writing styles. Copyists used different script types according to their place and date of activity (e.g. *praegothica*, *textualis*, *cursiva*, *semitextualis* [2]). Some script types were used preferentially according to the genre of the text under copy (e.g. liturgy, literature, diplomatic and pragmatic texts). Conversely, textual genres could influence some specific scribal practices (layout, abbreviations, etc.).

Pipeline description

Our aim is to evaluate the impact of data heterogeneity to build models for Latin and medieval French. Our corpus contains two levels of heterogeneity: it contains documents in one of two different languages (including internally some diatopic variation) [3], and variety of specifications for transcriptions. Each sentence of our corpus includes both abbreviated forms and expanded forms of words, thanks to the original encoding of the editions, that followed the Guidelines of the Text Encoding Initiative, and used a combination of <choice>, <abbr> and <expan> (TEI Consortium 2022).

Corpus have endured varying types of normalisations, sometimes contradictory (combined or decombined, etc.), to smooth discrepancies between transcriptions. The normalisation step follows this pipeline:

1. lowercasing;
2. normalising unicode (NFKD);
3. making substitutions based on an equivalence table and the use of "chocoMUFIn" (Clérice and Pinche 2021). In particular,
4.
   - normalisation of allographs (hypernormalised);
   - suppression of ramist distinctions (*u*/ *v* and *i*/ *j*);
   - removal of punctuation;
   - suppression of editorial marks (diacritics, apostrophes, cedillas, …).

We have divided our corpus into four training datasets to perform our evaluations and see potential benefits of fine-tuning for such an approach, on Latin or French texts and on abbreviated or expanded texts. The distribution of each corpus is described in table 2.

Table 2: Distribution of corpora into the four main datasets

	Abbr (Lines)	**Exp (Lines)**
	TOTAL : 8,528	**TOTAL :17,404**
	Fontenay : 1,365	Fontenay : 1,365
	MsDat : 2,217	MsDat : 2,217
	PsautierIMS : 3,086	PsautierIMS : 3,086
	StVictorLite : 1,860	StVictorFull : 10,736
	TOTAL : 19,532	**TOTAL : 19,530**
	ecmen : 9,831	ecmen : 9,831
	otinel bodmer : 1,977	otinel bodmer : 1,977
	otinel vaticane : 1,758	otinel vaticane : 1,758
	wauchier : 4,582	wauchier : 4,580
	Pelerinage : 1,384	Pelerinage : 1,384

Based on the experiments made by (Camps, Vidal-Gorène, and Vernet 2021) on abbreviated manuscripts, two approaches have been considered. Training on abbreviated data has been carried out with Kraken (Kiessling 2019; Kiessling et al. 2019), an OCR and HTR system previously used with success on a wide range of manuscripts (Camps, Clérice, and Pinche 2021; Scheithauer et al. 2021;

Thompson and others 2021), and training on expanded data with Calfa, an OCR and HTR system originally developped for highly abbreviated Oriental manuscripts (Vidal-Gorène et al. 2021). These two architectures use an encoder-decoder approach, the first one trained at the character level, the second one at the word level. If we keep the same hyperparameters defined previously (Camps, Vidal-Gorène, and Vernet 2021), we use a deeper architecture for the first one, architecture capable of high recognition rate in CREMMA (Pinche and Clérice 2021).

Preliminary results and discussion

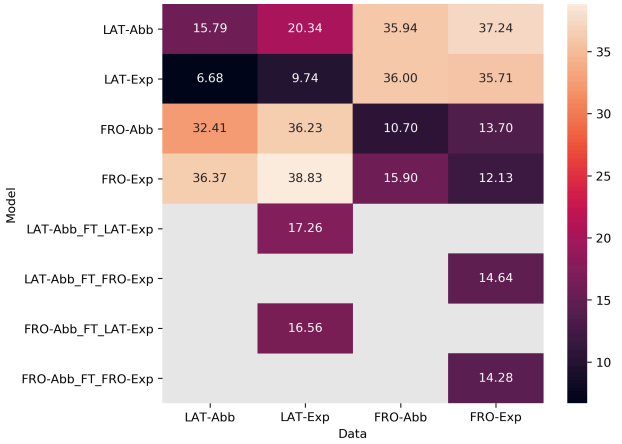
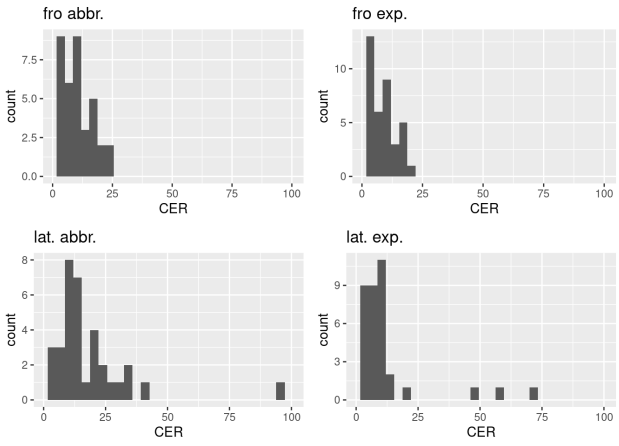


Figure 1:
Matrix of the cross evaluation of models



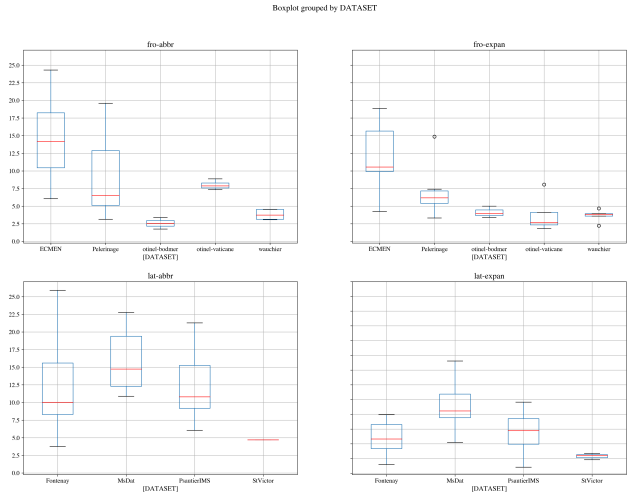


Figure 2:
Distribution of character error rates per page in the test sets; histograms (top) and boxplots (with outliers above 25% removed)

Current results show, perhaps counter intuitively, a better performance for expanded models, at least for Latin (fig. 1 and table 3), while, for French, the abbreviated model seem to perform slightly better (fig. 1). Perhaps more importantly, they show important variation in the distribution of the character error rates per page inside each test set and between test sets (fig. 3). Apart from a few strong outliers on the Latin corpora, with CER between 40 and 90% (due to issues in the test material), they show a situation that varies according mostly to the origin of the data. For some subcorpora, the CERs display very limited variation, with a very small interquartile range (CREMMA corpora for instance), while the results obtained for corpora such as ECMEN could reflect the larger variety of material they contain.

Nevertheless, among various observations, the following cases can be noted. On the one hand, on LAT-Exp predictions, the efficiency of the model is especially linked to the script used. Thus, the particularly angular *textualis quadrata*, widely used in PsautierIMS and some manuscripts of MSS dat. lat, is poorly recognised. We find a lot of issues related to the stems **ii** / **u** / **n** / etc. In the most extreme cases a significant difficulty in differentiating **c** and **e** occurs. For these scripts, tildes are seldom understood and abbreviations are therefore badly expanded. Meanwhile, in diplomatic texts of Fontenay, although the form of the letters is often sophisticated and flourished - especially in the first line of the charters - the model is able to recognise tildes and abbreviations. We also observe that the quality of the ink greatly influences the efficiency of the model. On the other hand, this multi-level heterogeneity seems to affect benefits we could expect of fine-tuning. We do not

notice any gain in recognition by fine-tuning abbreviated models with expanded data yet. Nevertheless we can already observe that cross-lingual fine-tuned models achieve similar recognition rates, even though abbreviations are widely different for these languages.

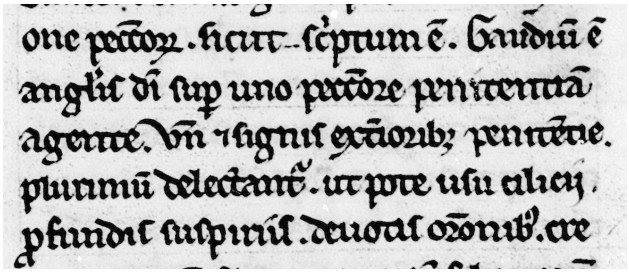


Table 3: Facsimile with ground truth abbreviated and extended and abbreviated and extended predictions from an extract of latin corpus of St Victor (BnF, Latin 14525, fol. 41va).

GT Abbr	GT Expan
one pecc̃oꝝ sicut sc̓ptum ẽ gaudiũ ẽangl̃is dĩ sup̱ uno pecc̃ore penitentiãagente uñ 7 signis ext̃iorib penitẽtieplurimũ delectant̃ ut pote usu ciliciiꝓfundis suspiriis deuotis or̃onib̃ cre	one peccatorum sicut scriptum est gaudium est angelis dei super uno peccatore penitentiamagente unde et signis exterioribus penitentieplurimum delectantur ut pote usu ciliciiprofundis suspiriis deuotis orationibus cre

Table 3: Facsimile with ground truth abbreviated and extended and abbreviated and extended predictions from an extract of latin corpus of St Victor.

Prediction Abbr	Prediction Expan
one **pecc̃oꝝꝝ** sicut sc̓ptum ẽ **sudiũ** ẽ angl̃is dĩ sup̱ uno pecc̃ore penitentiã agente uñ 7 signis ext̃iorib penitẽtie plurimũ delectant̃ ut pote usu cilicii ꝓfundis **suspinis** deuotis or̃onib̃ cre	one peccatorum sicut scriptum est **saudium** est angelis dei super uno peccatore penitentiam agente unde et signis exterioribus penitentie plurimum delectantur ut pote usu cilicii profundis suspiriis deuotis orationibus cre

All of this is deserving of further investigations, particularly to evaluate the impact of training set size versus training set diversity, and to measure the robustness of models trained with and applied to mixed language corpora. . Moreover, further normalisation of the training sets, and a direct inspection of outliers could allow to increase performance and intelligibility of the results.

Bibliography

Camps, Jean-Baptiste, Thibault Clérice, and Ariane Pinche. 2021. "Noisy Medieval Data, from Digitized Manuscript to Stylometric Analysis: Evaluating Paul Meyer's Hagiographic Hypothesis." *Digital Scholarship in the Humanities* 36 (Supplement_2): ii49–ii71. https://doi.org/10.1093/llc/fqab033.

Camps, Jean-Baptiste, Chahan Vidal-Gorène, and Marguerite Vernet. 2021. "Handling Heavily Abbreviated Manuscripts: HTR Engines Vs Text Normalisation Approaches." In *International Conference on Document Analysis and Recognition*, 306–16. Springer.

Clérice, Thibault, and Ariane Pinche. 2021. "Choco-Mufin, a tool for controlling characters used in OCR and HTR projects." https://doi.org/10.5281/zenodo.5356154.

Derolez, Albert. 2003. *The Palaeography of Gothic Manuscript Books from the Twelfth to the Early Sixteenth Century*. Cambridge Studies in Palaeography and Codicology 9. Cambridge: Cambridge University Press.

Kiessling, Benjamin. 2019. "Kraken - an Universal Text Recognizer for the Humanities." In *Proceedings of the Dh2019 Conference - Digital Humanities: Complexities, Utrecht, the Netherlands, 9–12 July 2019*. Utrecht: CLARIAH. https://dev.clariah.nl/files/dh2019/boa/0673.html.

Kiessling, Benjamin, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. 2019. "EScriptorium: An Open Source Platform for Historical Document Analysis." In *2019 International Conference on Document Analysis and Recognition Workshops (Icdarw)*, 2:19–19. IEEE.

Pinche, Ariane. 2021. "CREMMA Medieval, an Old French dataset for HTR and segmentation." https://doi.org/10.5281/zenodo.5235186.

Pinche, Ariane, and Thibault Clérice. 2021. "HTR-United/Cremma-Medieval: 1.0.1 Bicerin (Doi)." Zenodo. https://doi.org/10.5281/zenodo.5235186.

Scheithauer, Hugo, Alix Chagué, Rostaing Aurélia, Lucas Terriel, Laurent Romary, Marie-Françoise Limon-Bonnet, Benjamin Davy, et al. 2021. "Production d'un Modèle Affiné de Reconnaissance d'écriture Manuscrite Avec eScriptorium et évaluation de Ses Performances." In *Les Futurs Fantastiques-3e Conférence Internationale Sur L'Intelligence Artificielle Appliquée Aux Bibliothèques, Archives et Musées, Ai4lam*.

Stutzmann, Dominique, Christopher Tensmeyer, and Vincent Christlein. 2020. "Writer Identification and Script Classification. Two Tasks for a Common Understanding of Cultural Heritage." *Manuscript Cultures* 15: 11–24. https://www.csmc.uni-hamburg.de/publications/mc/files/articles/mc15-02-stutzmann.pdf.

Stuzmann, Dominique. 2011. "Paléographie Statistique Pour décrire, Identifier, Dater. . . Normaliser Pour Coopérer

et Aller Plus Loin ?" In *Kodikologie Und Paläographie Im Digitalen Zeitalter 2 - Codicology and Palaeography in the Digital Age 2*, edited by Franz Fischer, Christiane Fritze, and Georg Vogeler, 3:247–77. Norderstedt: Books on Demand (BoD).

TEI Consortium. 2022. "3.6.5 Abbreviations and Their Expansions." In *TEI P5: Guidelines for Electronic Text Encoding and Interchange, V4.4.0*. Text Encoding Initiative Consortium. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/CO.html#CONAAB.

Thompson, Walker, and others. 2021. "Using Handwritten Text Recognition (Htr) Tools to Transcribe Historical Multilingual Lexica." *Scripta & E-Scripta* 21: 217–31.

Vidal-Gorène, Chahan, Boris Dupin, Aliénor Decours-Perez, and Thomas Riccioli. 2021. "A Modular and Automated Annotation Platform for Handwritings: Evaluation on Under-Resourced Languages." In *International Conference on Document Analysis and Recognition*, 507–22. Springer.

Notes

1. We use the terminology graphemic (*graphématique*) and graphetic (*allographétique*) following (Stuzmann 2011).
2. For classification criteria and the lack of consensus among palaeographers, see (Derolez 2003; Stutzmann, Tensmeyer, and Christlein 2020).
3. We excluded documents with mixed contents (i.e., parts in French intertwined with parts in Latin), except for the ECMEN corpus which only contains small quotations or single words in Latin.

Distant reading of handwritten mid-nineteenth century Ottoman population registers

Can, Yekta Said
yektasaid.can@gmail.com
Koc University, Istanbul, Turkey

Kabadayi, M. Erdem
mkabadayi@ku.edu.tr
Koc University, Istanbul, Turkey

We propose an interdisciplinary paper in the fields of historical demography and computer vision based upon distant reading of mid-nineteenth century Ottoman population registers. Produced between the 1840s and the 1860s, these registers provide detailed demographic information on members of the households, i.e. names, family relations, ages, and occupations. The registers became available for research at the Ottoman state archives in Turkey, as recently as 2011. Their total number is around 11,000. Until now they have not been subject to any systematic study. Only individual registers were manually transliterated on a piecemeal manner. Source material for our case study consists of three population registers (NFS.d. 1865, 1866, 1867) belonging to the city of Manisa (See Fig 1).

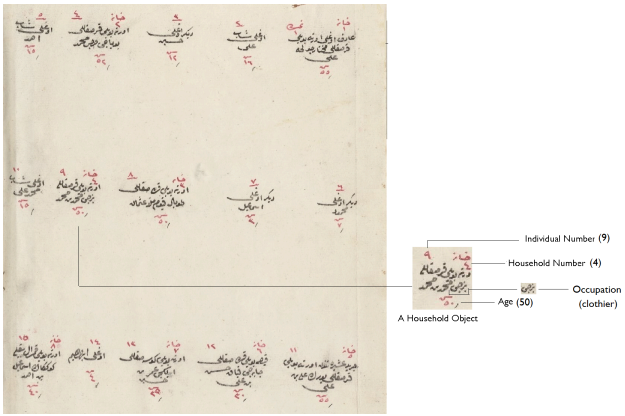


Figure 1:
*A sample register page and a household object are shown. The individual and household numbers are written at the top (line 1) of the individual data and ages are written in the last lines (line 4). Line 1: physical appearance (*uzun boylu kır sakallı*: tall, greybeard). Line 2: occupational title* bezci*: clothier), name (Mehmed son of Mehmed).*

In our ongoing multi-disciplinary research project, which has already started to contribute to digital and geospatial humanities with publications, we reached an unprecedented scale of manual data entry of Ottoman handwritten archival documents and data annotation possibilities for distant reading. Our project team located, read and entered data from a total of around 70 population registers belonging to around 100 locations. The total number and the order of individuals in registered on specific pages are also manually entered to our databases. Therefore, we could match automatically detected individual pixel positions in the page images to a large extent. From these registers our team of experts manually entered historical demographic information of a total of around 170,000 individuals living in around 70,000 households (see Fig. 2).

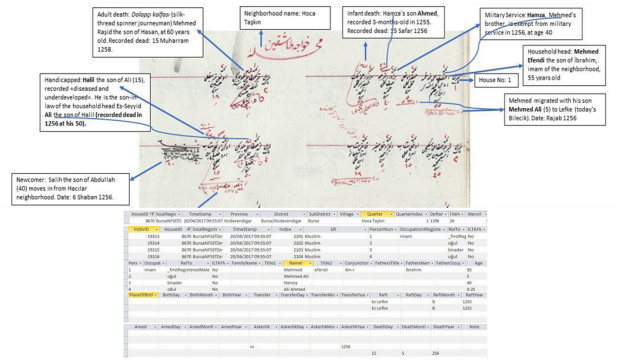


Figure 2:
Sample data entry and the manual annotation process for Ottoman population registers from Bursa.

We use this manually entered ground truth data to assess the effectiveness of our distant reading methods and calculate exact error rates. Our aim is to develop efficient distant reading methods based upon state-of-art computer vision technologies. Our work can be summarized in three steps. First, we segmented demographic information for each individual and, auto-counted persons in the city of Manisa. Then, we created a public Arabic handwritten digit dataset by manually spotting and cropping a selection of digits from the registers. Afterward, we automatically spotted the digits in our registers and recognized them using this dataset as training data. Lastly, we manually annotated occupations pixel-wise from the registers and created another publicly available dataset, comprising 400 handwritten occupational titles of 40 different occupations. We then spotted and identified the handwritten occupations automatically by using deep transfer learning (DTL). To the best of our knowledge, this is the first study to apply DTL on Arabic keyword and digit spotting.

With this paper we want to present not only our computing achievements to the DH community, but also problems we encountered in the process. Furthermore, we would like to underline implications of our new methodology in conducting research in digital history, historical demography and economic history.

The steps of our methodology: First, we developed a deep learning algorithm for detecting individuals in the population registers. We created a manually labeled dataset by using several registers and trained CNN models by using the dhSegment tool (Fig. 3, [Oliveira, SA, Benoit S,and Frederic K. "dhSegment: A generic deep-learning approach for document segmentation." 2018 16th ICFHR, IEEE]).

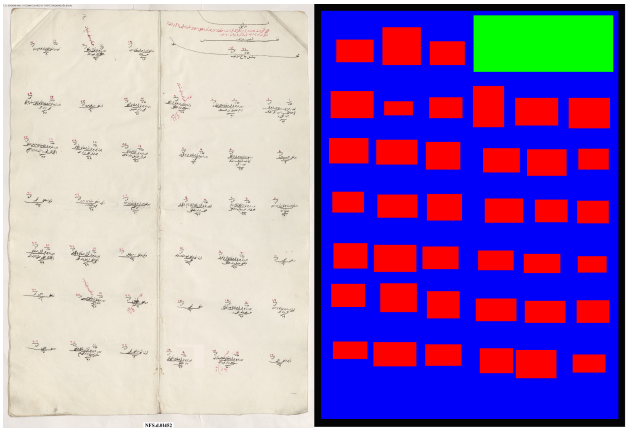


Figure 3:
The original page of a register is shown on the left side. The annotated label image is shown on the right side.

We trained different models for different types of layouts. The first model was trained for registers with tightly placed individuals. The second model was trained for registers with loosely placed individuals. We used the former model for Manisa registers of this study. After we detected the individual objects in these registers, by using the pixelwise locations, we cropped the demographic data of individuals to be used for line detection algorithms (Fig. 4).

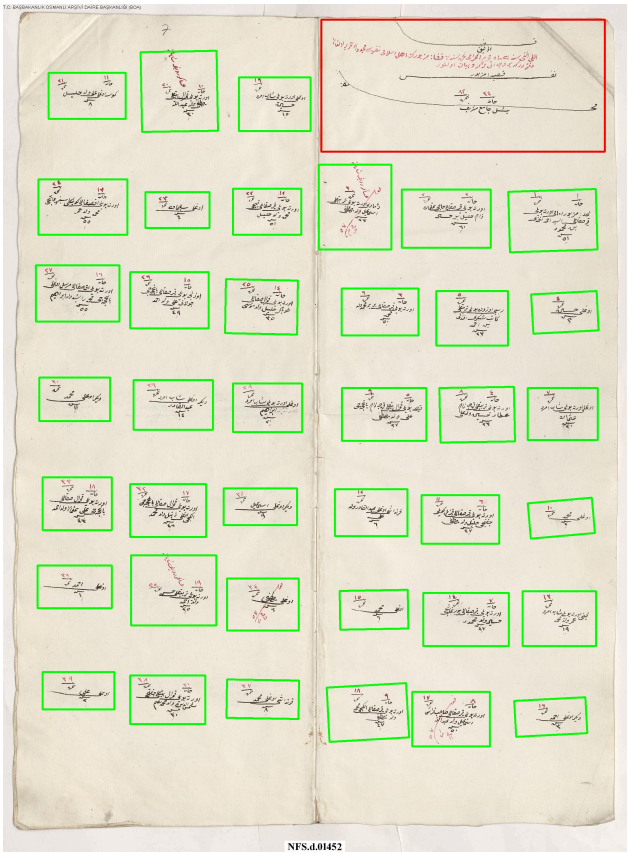


Figure 4:
Automatically detected individuals on a sample page.

In the Ottoman population registers series, the ages are written in the last line of each entry. Therefore, we applied a peak detection algorithm for detecting the last peak to separate the age information in the last line. As we are searching for the last line, we crop everything under the last and largest peak in the reverse HPP, which provides us the age of individuals (Fig. 5).

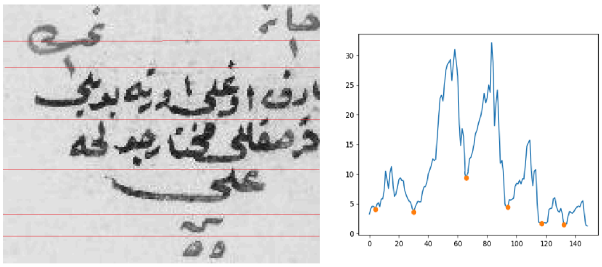


Figure 5:
A sample segmented individual in the left and corresponding valleys in HPP in the right are shown.

After line segmentation, we detected the digits, recognized them and directly assigned the digit value if there is only one digit. If there are two digits, we combined the digits into two-digit numbers by making necessary calculations. After predicting the ages, we retrieved the ground truth age of individuals from our databases, and then we draw the histogram of ages and compared it with ground truth distribution (Fig. 6).

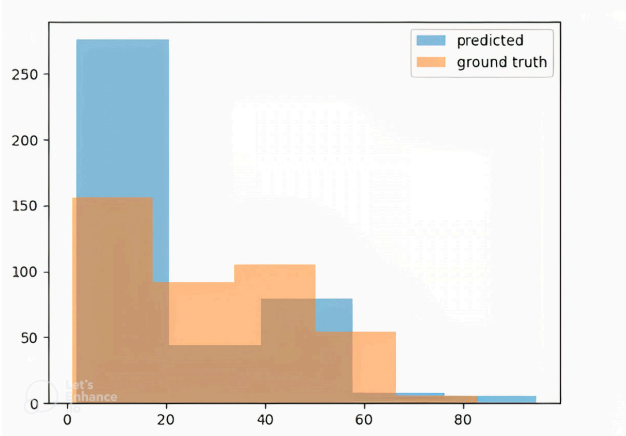


Figure 6:
The age distributions of ground truth ages and predictions are demonstrated.

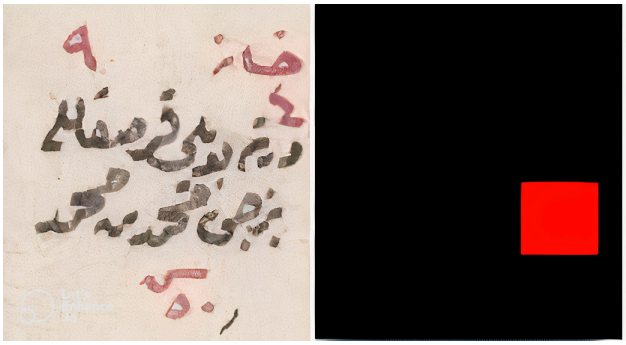


Figure 7:
An original cropped individual and the manually labeled occupational title are shown.

Lastly, we tried to identify occupational titles by creating an annotated public dataset including 400 handwritten occupational titles of 40 types (Fig. 7). To identify the occupational titles by using QbE (Query-by-Example), we applied DTL-based method in which we first trained a CNN network and used it to extract features from a pretrained Resnet-50 architecture and then identified the query images by applying k-nearest neighbor (kNN) algorithm on these features. In total, we extracted valuable information about number of people, age and occupational title information from these historical registers.

Implications for historical demography and economic history:

Our multi-stepped procedure can count total number of registered persons in given locations to a very high level of accuracy. The Ottoman population registers claims to cover entire male populace in the places they were conducted. The scribes did not reliably count the registered persons. Just to be able to automatically count total numbers of registered persons has huge potential for historical population studies using Ottoman archival documentation. Furthermore, we can now also count total number of households automatically. This feature has exploratory value for assessing household sizes for large parts of the mid-nineteenth century Ottoman Empire. Lastly distant reading of ages of individuals opens untapped possibilities for historical demography studies for around a dozen countries, territories of which were under the Ottoman rule in the mid-nineteenth century. Typically, several such countries such as Bulgaria, Greece and Serbia conducted their first population censuses after their independence and historical demography of these countries cannot reach back to pre-independence Ottoman era. Hence, making Ottoman population registers available on large scale via distant reading would revolutionize historical demography of not only Turkey, but also substantial part of Southeast Europe. Finally, sectoral coding of occupations and calculating their concentrations among major sectors (i.e. agriculture, industry, and services) provide economic historians with valuable proxy data to study the level of economic development. Therefore, methods of distant reading of occupations in large scale, such as are working on currently, will allow new in-depth studies on structural change of employment in the field of economic history again for substantial regions covered by the mid-nineteenth century Ottoman population registers.

Bibliography

Oliveira, SA, Benoit S,and Frederic K. "dhSegment: A generic deep-learning approach for document segmentation." 2018 16th ICFHR, IEEE

Named Entity Disambiguation for the Qing Shilu without Manually Labeled Data

Chao, Jo-Yu
v17291729@gmail.com
Department of Computer Science and Information Engineering, National Central University, Taiwan

Huang, Shi-Yun
as0210097@gate.sinica.edu.tw
Center for GIS, Research Center for Humanities and Social Sciences, Academia Sinica, Taiwan

Hsieh, Hsin-Yi
hsinmosyi@gmail.com
Department of Computer Science and Information Engineering, National Central University, Taiwan

Tsai, Richard Tzong-Han
thtsai@csie.ncu.edu.tw
Center for GIS, Research Center for Humanities and Social Sciences, Academia Sinica, Taiwan ; Department of Computer Science and Information Engineering, National Central University, Taiwan

Introduction

The study of historical figures is of great significance in the field of history. To investigate historical figures with digital humanity methods, the first step is to identify the names of people in texts. Not only is the person's name recognized from the text, but the person's name has to be linked to a knowledge base for reference. This is because the same person's name may refer to different people or other entities. This task is called named entity disambiguation (NED). The problem of historical figures with the same name is particularly serious when studying Manchu and Mongolian historical figures in the Qing Dynasty. A typical example is that the main persons (Cherin-Dorji, Yung-Te, Siang-Lin and Gui-Xiang) involved in the disaster report of the Kharkha Four Leagues (Ting-ting, 2016)have the same name as other historical figures. Fig. 1 shows that there are 10 Siang-Lins in the Ming-Qing Archives Name Authority Database (MQANAD).

德宗景皇帝實錄/卷之二百四十八/光緒十三年十月二十一日 Record of Emperor De Zong Jing/Vol. 248/October 21th of the year of Guang-Xu ○烏里雅蘇台參贊大臣祥麟等奏。 Siang-Lin (the Grand Minister Consultant of Uliastai) et al. report….
Y NO007969 祥麟 Siang-Lin
N NO007286 祥麟(Siang-Lin)
N NO007287 祥麟(Siang-Lin)
N NO007288 祥麟(Siang-Lin)
N NO007290 祥麟(Siang-Lin)
N NO007292 祥麟(Siang-Lin)
N NO007293 祥麟(Siang-Lin)
N NO007294 祥麟(Siang-Lin)
N NO007295 祥麟(Siang-Lin)
N NO007296 祥麟(Siang-Lin)

Fig. 1
There are ten people named Siang-Lin in the MQANAD database. Siang-Lin in this text refers to this historical figure with ID 0079695

Therefore, we select *Qing Shi-Lu*(*QSL*) and MQANAD(Liu, 2015) as the target text and the reference knowledge base for developing our NED system, respectively. *QSL* is the imperial annals of the Qing emperors, with a total of about 58 million characters, written in classical Chinese.

In (Tsai et al., 2020), several procedures were proposed to automatically generate labeled data from *Ming Shi-Lu* (*MSL*) for training a NED model. In this work, we improve (Tsai et al., 2020) in two folds. First, we propose a new way of expressing person names in both texts and profiles. Second, we modify the original procedures to improve the quality of the generated labeling data.

Method

As mentioned earlier, an NED (Cheng et al., 2019;Huang et al., 2015) model can link each mention to the correct profile in the reference knowledge base. An official's profile is shown in Table 1, we can see that it contains the official's attributes such as name, birth/death year, biography, resume, relatives, etc. The resume field is composed of all titles the official had held. We search *QSL* for all mentions of these officials' names. These mentions with the paragraphs containing them are used as our dataset. Table 2 shows an example instance of the search results.

ID	NO000033
Name	王安國 Wang An-Kuo
Gender	Male
Dynasty	Qing
Birth-Death year in Chinese calendar	康熙 33 年-乾隆 22 年 33rd year of Kang-xi to 22nd year of Qian-long
Birth-Death year in western calendar	1694 – 1757
Birthplace	江蘇省-揚州府-高郵州 (Gaoyou State, Yangzhou Prefecture, Jiangsu Province)
Relationship	王開運-曾祖父;王式<u>都</u>-祖父;王曾祿-父;王念孫-子;王引之-孫 Wang kai yun-Great grandfather; Wang shih ssu-grandfather; Wang tseng lu-father; Wang nien sun-son; Wang yin shih-grandson
Resume	都察院左僉都御史 the Left Assistant Censor-in-chief

Table 1
An official Wang An-Kuo's profile

ID	Name	Year	Book	Emperor	Volume	Year-Month	Text
NO000033	王安國 Wang An-Kuo	1739	清實錄 Qing Shilu	高宗 Kao Tsung	92	乾隆四年五月	以都察院左僉都御史王安國為左副都御史 Wang An-Kuo, the Left Assistant Censor-in-chief was appointed as the Left Vice Censor-in-chief.

Table 2
An example instance

We formulate this NED task as a text classification problem. Given an instance whose name field is m and a profile whose name field is also m, classify them as positive (matched) or negative (not matched). Matched means the person mentioned in the instance's paragraph field is just the official in the profile.

Tsai et al. was the first work to propose the procedures that compare instances and profiles to automatically generate labeled data to train the model (Tsai et al., 2020). In this work, we use all the procedures in (Tsai et al., 2020). These procedures based on the rules help us identify part of the data pairs. Then we deal with those data pairs that cannot be identified by the rules by using BERT as a binary classifier. With training data acquired from previous procedures, we make the classifier learn the relationships between instances and profiles and identify them through the context besides using rule-based procedures. In addition, we also propose methods to improve the quality of the labeled data, as described in the following paragraphs.

First, we replace the officials' names mentioned in personal profiles and instances with the symbol [unused_token], as shown in Fig. 2. This allows our model to focus on more information from the context and can be more robust to various person names.

德宗景皇帝實錄 卷之十一 光緒元年六月上 十二日
○以札薩克和碩親王車林多爾濟為烏里雅蘇台蒙古參贊大臣。

Record of Emperor De zong jing Vol. 11 June 12th of the 1th year of Guang-Xu
○Take Cherin-Dorji, Jasagh Imperial Prince, as Grand Minister Consultant of Uliasutai Mongolia.

Fig. 2
Cherin-Dorji, the official's name mentioned in personal profiles and instances, is replaced with the symbol [unused_token] (Hucker, 2008, Zhang et al., 2017)

Second, we have found many examples in the text where a person's name is the same as a location name. For example, Guilin (桂林) can be the name of a city or a minister's name. Therefore, we use a self-developed NER (Lin et al., 2020) system to process the texts. We use Flair as the NER model for this paper, the training data is from *MSL* and the test data is *QSL*'s 50 manually annotated data with F1 score 0.81. We revise (Tsai et al., 2020)'s method to correct an instance from positive to negative, an example is shown in Fig. 3.

N-n NO000852
桂林/世祖章皇帝實錄/卷之五十三/順治八年二月二十四日
Guilin /Record of Emperor Shi Zu Zhang/Vol. 53/February 24th of the 8th year of Shun-Zhi

○命靖南王耿繼茂。駐廣西(LOC)桂林(LOC)府
Command Geng Ji-Mao, the Prince of Jingnan to station in Guilin (LOC), Guangxi (LOC) city

Fig. 3
Guilin (桂林) is considered a location name by our NER system, therefore, it is generated as a negative instance

Lastly, since the text processed by (Tsai et al., 2020) is the Ming Dynasty, the text processed by this study is the Qing Dynasty, and the contents of *MSL* and *QSL* are slightly different, we only retain the first condition of Procedure 1.

Procedure 1: (our modified version)
For each profile f in MQANAD and each title t in f's resume field, we label the instances that have the same name as f's and satisfy either of the following two conditions as positive:
1. In the paragraph, t immediately followed by f's name
2. ~~f's name followed by any of characters in {為, 充, 兼, 調, 遷, 授} (all mean "work as") followed by t~~

Fig. 4
Procedure one is modified for writing-style difference between MSL *and* QSL

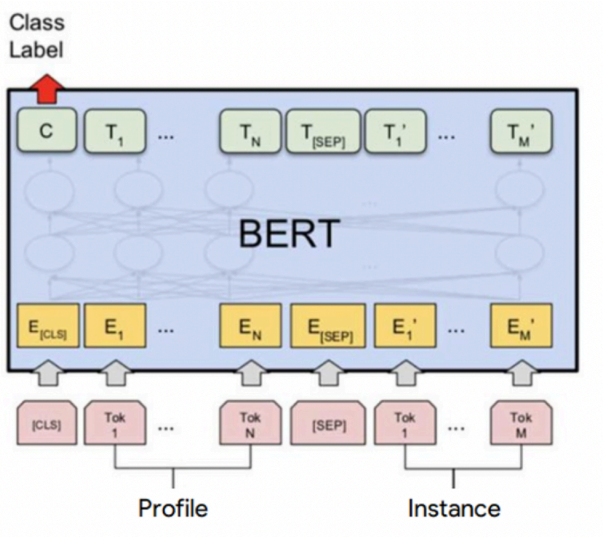


Fig. 5
Our BERT-based NED model

After automatically labeling the training data, we use BERT (Vaswani et al., 2017, Devlin et al., 2019), the state-of-the-art natural language understanding model, to perform text classification. We use the model pre-trained on the Chinese Wikipedia. As shown in Fig. 5, for each pair of profile and instance, we concatenate the cls symbol, the profile, the sep symbol, the instance as the input. The BERT model will output all label probabilities at the output position corresponding to the cls symbol.

We use the manually labeled data set of (Tsai et al., 2020) for performance evaluation. The results show that our NED model achieves an accuracy of 90%, which is 16% higher than the model proposed in (Tsai et al., 2020).

Finally, we conduct an ablation study. If we remove NER, performance will drop by 3%. If we do not do unused token replacement, performance will drop by 13%. If we do neither, the performance drops by 16%.

Conclusion

We have refined the approach of automatically generating labeled data for training an NED model. To be more specific, we employ our own NER system to eliminate the location names incorrectly labeled as positive instances and use the unused token symbol to enhance the robustness of our model. Results show that our refinement can improve the performance by 16% and our NED model will help us investigate historical figures in the Qing dynasty.

Bibliography

Cheng, J. et al. (2019). Entity Linking for Chinese Short Texts Based on BERT and Entity Name Embeddings.

Devlin, J. et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs].

Hucker, C.O. (2008) A Dictionary of Official Titles in Imperial China. Peking University Press.

Huang, H., Heck, L. and Ji, H. (2015). Leveraging Deep Neural Networks and Knowledge Graphs for Entity Disambiguation'. arXiv:1504.07678 [cs].

Institute of History and Philology, Academia Sinica (1984) Scripta Sinica. http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm.

Liu, C. (2015) Ming-Qing Archives Name Authority Database.

Ting-ting, L. (2016). 清季蒙古親王車林多爾濟史事鉤沉—以喀爾喀四盟報災案為中心[History of Mongolian Prince Cherin-Dorji in the Qing Dynasty-A Disaster Report by the Four Leagues of Khalkha]. Studies of Chinese Frontier Ethnic Groups.

Tsai, R.T.-H. et al. (2020). Automatic Labeled Data Generation for Person Named Entity Disambiguation on the Ming Shilu. DH.

Vaswani, A. et al. (2017). Attention Is All You Need. arXiv:1706.03762 [cs].

Lin, B.Y. et al. (2020). TriggerNER: Learning with Entity Triggers as Explanations for Named Entity Recognition. arXiv:2004.07493 [cs].

Zhang, Y. et al. (2017). 明代職官中英辭典 Chinese-English Dictionary of Ming Government Official Titles.

Digital Narrative of Buddhist Cave Temples: A Case Study of Niche 28 of Huangze Temple in Guangyuan, Sichuan

Chen, Wu Wei
wc54@nyu.edu
NYU Shanghai, China, People's Republic of

Zuo, Lala
lz2488@nyu.edu
NYU Shanghai, China, People's Republic of

The study of Buddhist cave temples is always an interdisciplinary effort. Art historians discuss artistic presentations through different media objects—— a Buddhist image could be depicted in a two-dimensional painting or a three-dimensional sculpture, or something in-between. Architectural historians explore the designs of the space and architectural features crossing different materials—— a cave temple could be an imitation of a wooden structure or blended with actual wooden parts. Other historians or scholars of religious studies explore social, cultural, or religious meanings from visual or textual data they find in the cave. It is not easy for people from different disciplines to communicate their knowledge and ideas even about the same cave. Not to mention the fact that passing such information to public visitors is even more difficult. Research institutes such as the Dunhuang Academy and the recently established Yungang Academy are demonstrations of such an interdisciplinary endeavor and how people are trying to break the disciplinary barriers. Now entering the digital age, we are equipped with new technologies such as cyber-archiving, A.I. analysis, GIS mapping, Extended Reality, among others. We may create a new form of narrative that helps people understand and appreciate the Buddhist cave temples, and more importantly, enable scholars to communicate more effectively across the times and disciplines. Therefore, this paper aims to identify the challenges, discuss possible solutions, and then use a case study to demonstrate our theories.

We will begin by discussing four challenges that digital methods may help to overcome. As mentioned earlier, a cave temple is a structure that contains enormous data about art, architecture, religion, society, and culture through visual (two-dimensional and three-dimensional), textual, and spatial forms. The first challenge is to integrate multi-layer information into one infrastructure. Secondly, it is still debatable if Buddhist cave temples are by definition "architecture." We may not consider a cave a "construction" but a "subtraction" from a mountain. The cave temples are often not completed at once, so the construction timelines are more complicated than most types of architecture. With optical and digital techniques, can we utilize them to tackle the problem? The third challenge is related to the immobility of the cave temples, while most stylistic analyses and comparisons rely on references to other sites. How can we virtually move the caves and visualize the comparison through the intermediate platform? Lastly, the conservation of the Buddhist caves is a high priority. The digital technology has been adapted in the 5-year plans over the last 20 years in China to document, compare, and prevent diseases, humidity, meanwhile monitoring micro weather, plate movement, and more to sustain the lifespan of "tangible heritage objects."

Following the theoretical discussion, we use the Niche 28 at Huangze Temple as a case study to demonstrate how to solve several problems from above. Niche 28,

also known as the Big Buddha Niche, is the largest-sized niche at Huangze Temple located in Guangyuan, Sichuan in southwest China. It is about 6.8 meters high, 5.5 meters wide, and 3.6 meters deep. Currently dated to the Sui dynasty (581-618AD), it is well-known for its representation of the Pure Land Buddhism paradise, where the five-meter tall Amitaba sculpture is flanked by two disciples, two bodhisattvas, and two guardians. The niche is also featured for the relief sculptures of eight classes of brave divine beings (*tianlong babu* 天龙八部) behind the sculptures. The structure is complicated, with the accessory niches flanking the main niche. Scholars also discover Tantric influence in iconography . Therefore, our first task is to present the complicated structure and enormous details using cyber-archiving. Digital documentation onsite enables to transform the physical, immovable sites into collaborative, interactive virtual sites online, and disseminate the research results accordingly. Through the GIS mapping, multi-layer of cultural heritage information (conservation, iconography, etc.) get to integrate cross-disciplinary perspectives towards the physical sites. Curatorial utilization is feasible through the overlaying of 3D tiles, geospatial analysis, photogrammetry texturing, and 4D timeline in the exhibitions or museum contents.

Another experiment is to verify different hypotheses about the construction date. Since no inscriptions have been found in this niche to prove the construction time directly, most people accept Wang Jianping's idea that the niche was constructed during the Sui dynasty while contesting theories still exist. Moreover, Wang relies on indirect evidence in Buddhist literature and stylistic comparisons with other sites. We hope to visualize Wang's comparison and argument to verify his hypothesis. In the traditional research routes, inscriptions and literature reviews provide convincing evidence. In the case of Niche 28, since the evidence is indirect, juxtapositions of digitized caves with similar features through the intermediate platform may provide new leads to the construction date.

Digital documentation, A.I. analysis, GIS mapping, Mixed Reality, and crowdsourcing will also apply to the conservation of the Big Buddha Niche. Digitization enables the lines to get further visible, and digital photo archives compare the migrating or unearthed statutes, some of which might have been relocated due to typhoons. With the GIS mapping, we can easily visualize the migration routes and the geo-locations. The site gets documented by integrated scanning. Raw data of the documentations are analyzed and re-established by SfM (Structure for Motion) technique. Point cloud data (.asc/.xyz/.ply) as the initial results can be further interpreted by modeling, texturing, and rendering to identify conservation concerns.

To conclude, we hope to explore a new model that can use digital technologies to facilitate the interdisciplinary approach in studying Buddhist cave temples—scholars are equipped with more tools to generate more research questions and testify their theories. A remote research platform and the exchange of digitized cultural assets for academic research are especially crucial during the Pandemic era. With substantial financial support, we will build an open-access platform with an open-source license. All scholars should have access to the Github repository.

A Novel Semi-supervised Framework to Identify Military Documents: A Quantitative Analysis on Military Records in *Ming Shi-Lu*

Chen, You-Jun
youjun1109@g.ucla.edu
Center for GIS, Research Center for Humanities and Social Sciences, Academia Sinica, Taiwan; Department of Mathematics, University of California, Los Angeles, CA, USA

Hsieh, Hsin-Yi
hsinmosyi@gmail.com
Center for GIS, Research Center for Humanities and Social Sciences, Academia Sinica, Taiwan; Department of Computer Science and Information Engineering, National Central University, Taiwan

Tsai, Richard Tzong-Han
thtsai@g.ncu.edu.tw
Center for GIS, Research Center for Humanities and Social Sciences, Academia Sinica, Taiwan; Department of Computer Science and Information Engineering, National Central University, Taiwan

Introduction

Faced with an extensive digitized corpus to analyze, a historian may find text analysis or concordance software such as Antconc, utilizing context-independent and word-count approach, desirable for identifying relevant sentences or documents in which words of interests located. However,

a potential pitfall of such a methodology is that a particular keyword's existence or absence does not necessarily entail discussing a subject (Bingham, 2010), especially for those exploring specific phenomena or broad social or cultural themes.

This work presents a semi-supervised deep neural framework, leveraging contextualized representation learning techniques, to automatically identify military documents in Ming Shi-Lu, without previously labeled data. We aim to expedite the onerous document compilation process of relevant and consistent information about the phenomenon to be investigated. In particular, we leverage a weakly supervised classification model (WSM) built upon the study by (Meng et al., 2020), which emulates how humans categorize documents into named categories to generate high-confidence labeled data. The quantitative results in our analysis, contributing another dimension toward studying the Ming military, lend credence to the effectiveness of our approach and shed light on the development and collapse of the Ming empire.

Background

Ming Shi-Lu (*MSL*), composed of 208,522 records, is an official annalistic work centering on Ming emperors compiled by the officials in the Ming Dynasty (1368 A.D. to 1644 A.D.). Preserving enormous original documents of edicts, decree, and records of political, military, socio-economic, and other major events, *MSL* plays an essential role in the historical reconstruction of the diverse Eastern Asian societies and polities. Our study focuses on military records in *MSL,* aiming to unveil their underlying historical, academic, and documentary value by natural language processing techniques.

Methodologies

Even though the emergence of pre-trained language model (LM) has drastically reduced the amount of training data needed for supervised methods, experiments show that the amount of training data still has to reach 4-5% of the entire dataset (Grießhaber et al., 2020), which in our case approximates to 5,400 - 6,900 labeled documents, to yield steady and satisfying accuracy.

The proposed framework overcomes the need for previously labeled data (Fig. 1). The included steps are: (1) select a small set of seed words describing the categories to be classified, (2) use a WSM to produce labeled data according to the selected words, (3) rearrange the resulting labeled data into a training set, and (4) use a supervised classification model (SCM) to categorize the entire dataset.

Our WSM is based on **LOTClass** (Meng et al., 2020), which leverages the **BERT**-base-chinese language model (Devlin et al., 2019) as the general knowledge base for category name understanding and feature representation learning model for classification. The BERT-based-chinese model is also the backbone for our SCM.

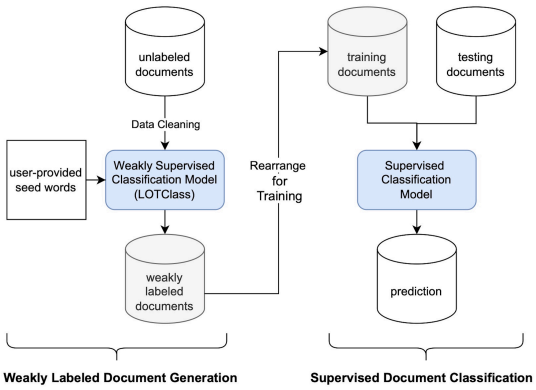


Fig. 1 The generic framework in this study . (1) The WSM will (i) learn a set of category indicative vocabulary from user-provided seed words of each class (Table 2 and 3), (ii) find the category-indicative words (w) in the text, and train the model via cross-entropy loss with a classifier on top of each w's contextualized embedding to predict their implied categories, (iii) generalize through a self-training mechanism, and (iv) make predictions (Table 3). (2) The process of supervised document classification includes: (i) take a set of documents from the prediction results of WSM as ground truth labels (ii) fine-tune a pre-trained LM on a classification task with the labeled data, (iii) evaluate the model, and (iv) use the trained model to predict the remaining set of the documents

Dataset and definition of military documents

We define the military documents as records containing offensive or defensive operations of both combat and non-combat nature and consider only documents involving human activities in *MSL (Institute of History and Philology, Academia Sinica, 1984) ,* which is in full 136,427 documents .

Weakly labeled data generation

We use trivial military action words such as attack and defense as seed words for military documents. To capture

non-military documents, we also select 20 different classes of seed words (Table 1). The WSM will generate a set of category indicative vocabulary (Table 2) based on the input seed words for each class and leverage the contextualized category indicative words to classify the documents (Table 3).

Table 1 User-provided seed words of Category Lawbreaking, Repair, and Military in our study . We carefully choose and expand univocal characters as seed words by inspecting the classification results of WSM

Class	Seed Words
Law-breaking	罪 (fault), 罰 (punish), 獄 (prison/jail), 隱 (conceal)
Repair	修 (fix/repair), 築 (build)
Military	率 (lead), 守 (guard/defend), 攻 (attack), 擊 (assault), 備 (guard/defend)

Table 2 Category vocabulary of Category Lawbreaking, Repair, and Military in our study

Class	Category-indicative Vocabulary
Law-breaking	罪 (fault/sin), 罰 (punish), 獄 (prison/jail), 辜 (crime/guilt/sin), 訴 (indict), 嫌 (disgust), 條 (ordinance), 惡 (evil), 違 (transgress), 錯 (fault), 讞 (conviction), 累 (fault), 論 (infer), 義 (justice), 律 (law), 冤 (injustice/wrong), 誅 (punish), 諒 (forgive), 訟 (litigation), 罚 (punish), 言(state), 狱 (prison/jail), 型 (canonical), 牢 (prison/jail)
Repair	修 (repair), 脩 (build), 築 (build), 繕 (mend), 填 (fill), 勘 (reconnaissance), 刪 (remove), 筑 (build), 恢 (restore), 維 (maintain), 闢 (reclamation), 茸 (thatch/repair), 完 (complete), 纂 (redact), 善 (complete), 裁 (cut), 刻 (carve), 圍 (enclose), 拆 (dismantle), 夯 (tamp), 織 (weave), 蓋 (build/construct), 砌 (build by laying bricks), 勤 (industrious), 堵 (block), 搭 (put up/build), 播 (sow), 閱 (inspect), 浚 (dredge), 竣 (finish), 程 (procedure), 彌 (fill), 辟 (open up), 匡 (correct)
Military	守 (guard/defend), 禦 (defend), 備 (defend), 攻 (attack), 御 (withstand), 準 (prepare), 擊 (assault), 候 (watch), 照 (look after/guard), 戒 (guard), 备(guard), 侍 (serve), 剿 (suppress), 略 (strategy), 預 (prepare), 標 (target), 揮 (wield)

Table 3 Examples of Category Lawbreaking, Repair, and Military prediction results by WSM. (1) Examples 1, 5, and 6 have no user-provided seed words or category indicative vocabulary. This implies that the model can identify a document's category without trivial keywords, surmounting the limitation of the keyword search approach. (2) We take 4,000 documents from Category Military and 4,000 from the rest of the non-military categories for training data

Number	Class	Prediction Results
1	Law-breaking	下近侍陳忠于法司究問以其連結外人陳槐同謀妄告也 Eunuch Attendant Zhong Chen was interrogated in court because he conspired with the outsider Huai Chen to make false accusations.
2	Law-breaking	南京戶科給事中甄成德劾奏南京刑部尚書顧璘不職詔璘回藉聽勘 Chengde Zhen, Supervising Secretary of the Office of Scrutiny for Revenue in Nanjing (南京戶科給事中), wrote a letter to the Emperor impeaching Lin Gu, Minister of Justice in Nanjing (南京刑部尚書), for malfeasance. The Emperor ordered Lin to return to his hometown and wait for an investigation.
3	Repair	雲南巡撫陳用賓等以蒙猓既平條議善後七事一建縣二建哨三勘田四寬賦五工費六哨役七擇官俱允行 After suppressing the barbarians Mu Pu (蒙猓), Yongbin Chen, Grand Coordinator (巡撫) of Yunnan, and other officials discussed with them about the seven things to cope with the aftermath of suppression: First, establish the county. Second, build sentries. Third, conduct a land investigation. Fourth, relieve taxes. Fifth, pay labor fees; Sixth, assign sentinels. Seventh, perform official selections. Mu Pu promised to implement all.
4	Repair	命錦衣衛指揮僉事宗鐸監察御史秦顯王璧巡視居庸山海紫荊等関修塞隘口開掘溝塹 Order Assistant Commander of Imperial Bodyguard (錦衣衛指揮僉事) Zong Duo, and Investigating Censor of the Zhejiang Circuit (監察御史) Yong Qin and Bi Wan, to patrol Juyong, Shanhai, Zijing, and other places, repair the passes and dig trenches.
5	Military	寧夏東路花馬池伏羌等墩寇入境報至勑總兵官楊信馳赴寧夏調兵應援 It is reported that enemies had invaded Hua Ma Chi and Fu Qiang on Ningxia East Road. Regional Commander (總兵官) Xin Yang was ordered to head to Ningxia and dispatch troops to rescue quickly.
6	Military	勑都督毛福壽等於京城外西南街巷要路堵塞路口埋伏神銳短銃以待策應 Order Commissioner-in-Chief (都督) Fushou Mao and others to block the intersections of the main streets outside the southwest of the capital and wait in ambush with sharp swords.

BERT-based supervised document classification

To initiate the supervised document classification task, we take 8,000 labeled documents generated by WSM. The examination result by manually examining 5% of the labeled documents shows that the WSM achieves around 87.3% accuracy. We then fine-tune the bert-base-chinese model for a binary classification task by randomly taking 80% of the labeled data for training and 20% for validation, and evaluating the model performance via precision, recall, and f1 scores (Table 4). Subsequently, we use the trained binary classifier to predict the rest of 128,427 documents.

Table 4 Precision, recall, and f1 scores of the BERT-based binary classification model

	Precision	Recall	F1
Training set	0.89	0.89	0.89
Validation set	0.90	0.90	0.90

Result and analysis

Comparison with the distribution of war frequency in the Ming Dynasty

As armed forces are primarily intended for warfare, we compare the distribution of the number of military documents with the distribution of war frequency (W) (Editorial Committee of Chinese Military History, 1985) in the Ming Dynasty (Fig. 2). It can be seen that, up to the official compilation of *MSL* (1627 A.D.), the fluctuations of both trends match, evincing the robustness of our framework.

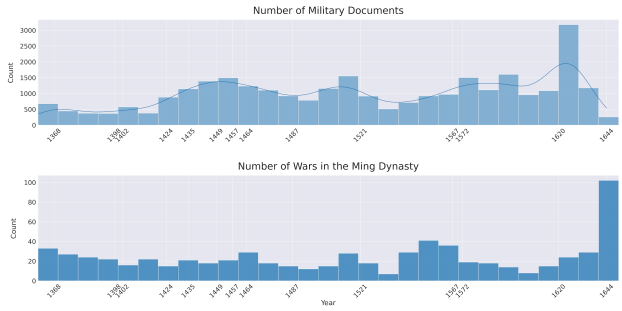


Fig. 2 Distributions of the Number of Military Documents in *Ming Shi-Lu* **and Number of Wars in the Ming Dynasty .**(1) The x-axis labels show the duration of reign for each Ming Emperor. (2) The divergence of the two trends after the 1630s may be explained by two factors: (i) The official *MSL* only covers reigns from the Hongwu Emperor (1368 A.D. - 1398 A.D.) to the Tianqi Emperor (1605 A.D. - 1627 A.D.). Records of the Chongzhen Emperor (1627 A.D. - 1644A.D.), the last Ming emperor, are from *Chongzhen Shi-Lu* and *Chongzhen Chang-Bian*placed in the appendix of *MSL*, even though providing an account of the reign, yet significantly fewer in numbers than records of other reigns. (ii)The last interval of the Chinese war data (W), which we have access to, ranges from 1640 to 1649, outrunning the rule of the Ming empire

Evaluation of military document ratio distribution in *Ming Shi-Lu*

To identify high-density periods of military documents, we convert the absolute number of military documents into ratio distribution calculated on a 5-year interval (Fig. 3).

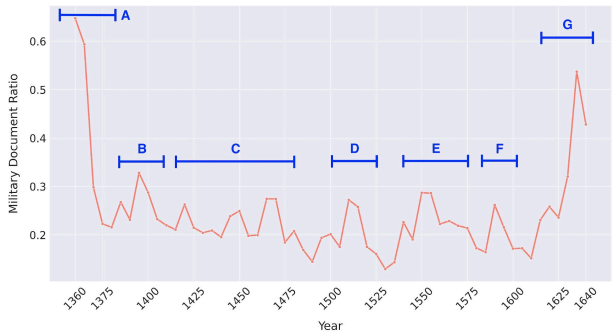


Fig. 3 Military Document Ratio Distribution in *Ming Shi-Lu* **.**We marked the peaks and listed the corresponding major military events below: (A) The founding of the Ming dynasty. (B) The Jignan campaign (1399-1402) and the Yongle Emperor's campaigns against the Mongols (1410-1424). (C) The War of Lucha (1436-1449), Tumu Crisis (1449), and the Defense of Jingshi (1449). (D) Wokou raids and Mongols raids. (E) The Jiajing wokou raids andthe War of Gengxu (1550). (F) The Bozhou campaign (1589-1600), the Ningxia campaign (1592), and the Imjin War (1592-1598). (G) The Battle of Sarhū and the collapse of the Ming dynasty. Fig. 2 (2) (i) can explain the anomalous decline in ratio during 1640-1644

The successive peaks (B - E) coincide with major military events of profound influence on the development of the Ming Dynasty. The Jignan campaign (B) attributed to a drastic change in the country's military defense system. The incredible feats achieved by the Ming in the War of Luchuan (C) indirectly influenced the origin of civil officials' exercising military power (Li, 2003). The War of Gengxu (E) occurred while Ming armies suffered repeated defeats in combating the Wokou and the Mongols raids on the Ming territory led to the abolishment of the Superintendent of the Integrated Division established for a century (Wu, 2021).

During the later reign of Wanli Emperor, the increased military expenditure and the exacerbation of the fiscal crisis resulting from the military campaigns (F) brought about the downfall of the Ming Dynasty (Zhao et al., 2016). Additionally, pleasantry uprisings instigated by severe droughts in 1627-1643 and the ensuing famine, and the southward migration of the Mongols caused by the effects of the Little Ice Age precipitated the collapse of the Ming Empire (Zheng et al., 2014; Sun and Zhang, 2018).

Conclusion

This work introduces a semi-supervised framework identifying military documents without any labeled data, significantly reducing the manual labeling effort by domain experts. Empirical results in our analysis, aligning with the

occurrence of major campaigns, demonstrate the robustness of our approach. For future work, we would like to continue exploring the potential of this framework and apply it to existing Asian corpora such as *Veritable Records of the Joseon Dynasty*, contributing to the reconstruction of diverse Asian history. Additionally, we plan to conduct an in-depth investigation on the military documents in *MSL* to substantiate perceived historical hypotheses with quantitative, temporal, or geographical evidence.

Bibliography

Bingham, A. (2010). 'The Digitization of Newspaper Archives: Opportunities and Challenges for Historians'. *Twentieth Century British History*, 21(2): 225–31 doi:10.1093/tcbh/hwq007.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–86 doi:10.18653/v1/N19-1423.

Editorial Committee of Chinese Military History (1985). Tabulation of Wars in Ancient China. *People's Liberation Army Press*, Beijin.

Grießhaber, D., Maucher, J. and Vu, N. T. (2020). Fine-tuning BERT for Low-Resource Natural Language Understanding via Active Learning. *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 1158–71 doi:10.18653/v1/2020.coling-main.100.

Institute of History and Philology, Academia Sinica (1984). Scripta Sinica http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm.

Li, F. (2003). 明代文人對軍隊的統領論析 [The Civil Officers' Command of the Army in the Ming Dynasty] Master Thesis. http://cnki.sris.com.tw/kns55/brief/result.aspx?dbPrefix=CMFD.

Meng, Y., Zhang, Y., Huang, J., Xiong, C., Ji, H., Zhang, C. and Han, J. (2020). Text Classification Using Label Names Only: A Language Model Self-Training Approach. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 9006–17 doi:10.18653/v1/2020.emnlp-main.724.

Sun, C. and Zhang, Q. (2018). 氣候變遷、政府能力與王朝興衰——基于中國兩千年來歷史經驗的實證研究 [Climate Change, State Capacity, and the Rise and Fall of Dynasties——An Empirical Study Based on Chinese Historical Experience in the Past 2000 Years]. *China Economic Quarterly,* 18(1): 311–36.

Wu, Y. (2021). 六師之任：明代協理京營戎政與北京防禦 [Governor-General of Capital Defenses and Military Affairs of Beijing Area in Ming China]. Ph.D. thesis, National Taiwan Normal University.

Zhao, Y., Fu, W. and Hei, W. (2016). 試析播州之役對明朝的影響 [The Influence of the Bozhou rebellion on the Ming Dynasty]. *Wenjiao Ziliao*, 19: 72–73.

Zheng, J., Xiao, L., Fang, X., Hao, Z., Ge, Q. and Li, B. (2014). How climate change impacted the collapse of the Ming dynasty. *Climatic Change*, 127(2): 169–82 doi:10.1007/s10584-014-1244-7.

Linked Open Dictionaries (2015-2022): Achievements, Experiences and Challenges with respect to LOD Technology in Linguistics and the Philologies

Chiarcos, Christian
christian.chiarcos@gmail.com
Applied Computational Linguistics, Goethe Universität Frankfurt, Germany

Ionov, Maxim
max.ionov@gmail.com
Applied Computational Linguistics, Goethe Universität Frankfurt, Germany

Fäth, Christian
faeth@em.uni-frankfurt.de
Applied Computational Linguistics, Goethe Universität Frankfurt, Germany

The project Linked Open Dictionaries (LiODi) was funded in the eHumanities programme of the German Federal Ministry for Education and Research (BMBF, 2015-2022) and conducted in collaboration between empirical linguistics and computational linguistics at the Goethe University Frankfurt, Germany.

Our goals were two-fold. In terms of humanities, it pursued the study of language contact in the Caucasus, especially on North-Eastern Caucasian and Armenian in their contact with Kartvelian (Georgian), Iranian and Turkic

(Rind-Pawlowski 2017, Chiarcos et al. 2018, Bellamy and Schreur 2021). In terms of digital methodology, the project pioneered, explored and advocated the use of RDF and Linked Data formalisms for research questions in the philologies and linguistic typology. We summarize this aspect, our achievements, experiments, and challenges encountered, and we discuss implications for the future use of Linked Data and RDF technologies in linguistics and philologies.

While Semantic Web technology has been well established in Digital Humanities prior to the project, this was largely restricted to prosopography, entity linking and object metadata; the potential of Linked Data to achieve interoperability for dictionaries and digital editions still remained underexplored. This situation changed, and to some extent, our project laid the groundwork for subsequent DH projects that adopted technologies and formalisms developed by the project.

The project initiated and contributed to community standards for various language resources:

- OntoLex: OntoLex-Lemon is a community standard for lexical resources. Inspired by the application of Monnet-Lemon to the conjoint development of ontologies and dictionaries (Weingart and Giovannetti 2016), LiODi engaged with the W3C Community Group Ontology-Lexica for developing novel OntoLex modules for morphology (OntoLex-Morph; Klimek et al. 2019) and corpus information (OntoLex-FrAC, Chiarcos et al. 2020b)
- Ligt: We developed an RDF vocabulary for interlinear glossed text, in order to overcome technological barriers between conventionally used software used for glossing (Chiarcos et al. 2017, Ionov 2021).
- CoNLL-RDF: We developed an RDF vocabulary and a converter suite for annotations as used in NLP and corpus linguistics (Chiarcos & Fäth 2017).
- TEI+RDFa: Following a survey of the relation of TEI and RDF (Chiarcos and Ionov 2019), we proposed RDFa for an inline XML representation. Together with the Academy of Sciences in Heidelberg, Germany, and the POSTDATA project, we provided the first implementation of TEI+RDFa for a small-scale digital edition and demonstrated its benefits for linking and querying via open web services (Tittel et al. 2018).

All our code and all distributable data are available over our GitHub repository (https://github.com/acoli-repo/, https://acoli-repo.github.io/liodi). Language resources produced on this basis include the single largest collection of machine-readable open source bi-dictionaries, the ACoLi Dictionary Graph, published in accordance with Linked Data principles and using OntoLex-Lemon (Chiarcos et al. 2020a, see Fig. 1). In terms of software, notable contributions include tools for the detection of cognates and loan words and for translation inference across dictionaries (Chiarcos et al. 2020c).

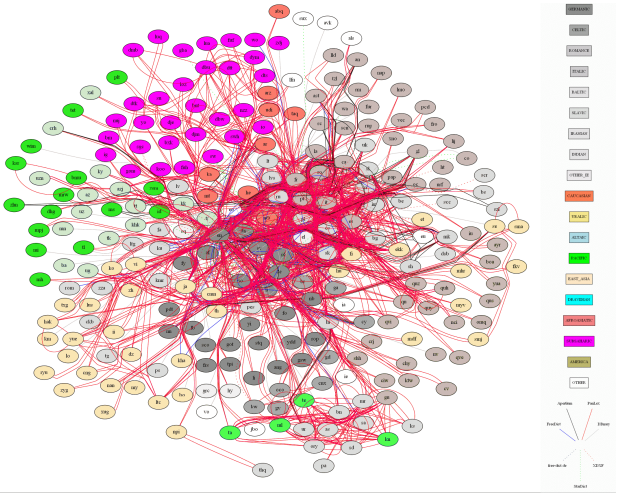


Fig. 1.
The ACoLi Dictionary graph: 3000+ bi-dictionaries (edges in the diagram) for more than 400 languages (nodes in the diagram) in machine-readable formats (OntoLex-Lemon) and published as open source, source: https://github.com/acoli-repo/acoli-dicts. Colors indicate language families or regions.

Furthermore, a main contribution of the project lay in documentation and dissemination of LOD technology in linguistics, language technology and philologies. On an international level, this includes the first text book on the topic (Cimiano et al. 2020), the organization of workshops and conferences, and four summer schools/datathons (SDLLOD 2017, 2019, 2022; EUROLAN 2021). In particular, our datathons proved highly effective in disseminating LOD expertise into other LOD-affine projects and contributed to the conception of novel projects by third parties. This includes the use of LOD in academic publishing (Nordhoff 2020, Ligt), infrastructure projects and community portals (Page-Perron et al. 2017 for cuneiform: CoNLL-RDF; Mambrini and Passarotti 2018, Pellegrini et al. 2021 for Latin: OntoLex-Morph, CoNLL-RDF) and projects for digital edition and lexicography (Mondaca and Rau 2020: OntoLex and TEI+RDFa; Chiarcos et al., accepted: TEI +RDFa).

As for key experiences, there are three main conclusions we arrived at at the end of the project:

- LOD and RDF technologies are ideally suited as middle-ware for DH projects. They facilitate information integration between and linking from

various resources. In particular, it is not required to abandon established workflows, but only an RDF wrapper for their components.
- Existing RDF technology is sufficiently versatile to provide such an additional layer of interoperability over existing solutions and workflows with moderate effort. W3C standards and RDF serializations provide interoperability with XML (RDFa), JSON (JSON-LD), tabular formats (RDB2RDF, CoNLL-RDF), etc.
- RDF technology is generic, that is, in many cases not sufficiently optimized for real-time performance. Although we provide technology running on RDF backends (e.g., CQP4RDF, Ionov et al. 2020), we see the main benefit of RDF technology at the moment in the backend and between existing tools rather than as their basis, mostly for reasons of backward-compatibility. In the end, we moved the focus of the project from providing a designated new toolbox into the development of a technology stack that allowed researchers to continue working with their existing tools, and then integrate and link the results.

Finally, we also conclude with a somewhat bitter note: The project produced considerable amounts of open source data, but we are now in a situation where we could not secure adequate long-term hosting. This is not so much because of the lack of hosting solutions, but that we found that none of the portals we explored would support depositing data dumps in a way that they provide resolvable URIs. So, we can (and do) make our open source RDF data available, but we cannot provide it as linked data, because academic solutions such as Zenodo and commercial providers such as GitHub do not allow to declare the data with the adequate mediatype but force data providers to resort to text/plain or application/octet-stream (Chiarcos 2021).

When trying to access and process this data, especially in federated search, tools and users may thus be confronted with their SPARQL engines producing unpredictable results, as the correct format needs to be guessed heuristically, and existing SPARQL engines differ in their behaviour. In our opinion, the lack of such hosting solutions (at least in Europe) contributes significantly to the perceived instability and unreliability of LOD-based technologies. This is, however, less a technical problem than a political or administrative one, and one we would like to discuss with the community.

Bibliography

Bellamy, K., and Schreur, J. W. (2019). Multiple Methods for Investigating Code-Switching in Batsbi Nominal Constructions. In *Linguistic Forum 2019: Indigeneous Languages of Russia and Beyond* (p. 20).

Chiarcos, C., and Fäth, C. (2017). CoNLL-RDF: Linked corpora done in an NLP-friendly way. In *International Conference on Language, Data and Knowledge* (pp. 74-88). Springer, Cham.

Chiarcos, C. and Ionov, M. (2019). Linking the TEI: Approaches, Limitations, Use Cases. In *Digital Humanities Conference 2019 (DH2019)*, Utrecht University, July.

Chiarcos, C., Ionov, M., Rind-Pawlowski, M., Fäth, C., Schreur, J. W., and Nevskaya, I. (2017). LLODifying linguistic glosses. In *International Conference on Language, Data and Knowledge* (pp. 89-103). Springer, Cham.

Chiarcos, C., Donandt, K., Sargsian, H., Ionov, M., and Schreur, J. W. (2018). Towards LLOD-based language contact studies. A case study in interoperability. In *Proceedings of LREC-2018*, May 2018, Miyazaki, Japan.

Chiarcos, C., Fäth, C., and Ionov, M. (2020a). The ACoLi dictionary graph. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 3281-3290).

Chiarcos, C., Ionov, M., de Does, J., Depuydt, K., Khan, F., Stolk, S., Declerck, T., and McCrae, J. P. (2020b, May). Modelling frequency and attestations for Ontolex-lemon. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography* (pp. 1-9).

Chiarcos, C., Schenk, N., and Fäth, C. (2020c). Translation Inference by Concept Propagation. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography* (pp. 98-105).

Chiarcos, C. (2021). Get! Mimetypes! Right!. In *Proceedings of the 3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Chiarcos, C., Gelumbeckaite, J., Drach, M., and Ionov, M. (accepted). *The Postil Time Machine: The Lithuanian Lutheran Postils of the 16th Century*, accepted at DH2022.

Cimiano, P., Chiarcos, C., McCrae, J. P., and Gracia, J. (2020). *Linguistic Linked Data*. Springer International Publishing.

Ionov, M. (2021). APiCS-Ligt: Towards Semantic Enrichment of Interlinear Glossed Text. In *P roceedings of the 3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Ionov, M., Stein, F., Sehgal, S., & Chiarcos, C. (2020). cqp4rdf: Towards a Suite for RDF-Based Corpus Linguistics. In *Proceedings of the European Semantic Web Conference (ESWC-2020)*. Springer, Cham.

Klimek, B., McCrae, J. P., Bosque-Gil, J., Ionov, M., Tauber, J. K., and Chiarcos, C. (2019). Challenges for the representation of morphology in ontology lexicons. In *Proceedings of eLex- 2019*.

Mambrini, F., and Passarotti, M. (2019). Linked Open Treebanks. Interlinking Syntactically Annotated Corpora in the LiLa Knowledge Base of Linguistic Resources for Latin. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)* (pp. 74-81).

Mondaca, F., and Rau, F. (2020). Transforming the Cologne Digital Sanskrit Dictionaries into Ontolex-Lemon. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)* (pp. 11-14).

Nordhoff, S. (2020). Modelling and annotating interlinear glossed text from 280 different endangered languages as Linked Data with LIGT. In *Proceedings of the 14th Linguistic Annotation Workshop* (pp. 93-104).

Pagé-Perron, E., Sukhareva, M., Khait, I., Chiarcos, C. (2017). Machine translation and automated analysis of the Sumerian language. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 10-16).

Pellegrini, M., Litta Modignani Picozzi, E. M. G., Passarotti, M., Mambrini, F., and Moretti, G. (2021). The Two Approaches to Word Formation in the LiLa Knowledge Base of Latin Resources. In *Proceedings of the Third International Workshop on Resources and Tools for Derivational Morphology* (pp. 101-109). ATILF.

Rind-Pawlowski, M. (2017). Formation and function of the imperfective stems in Khinalug. In *Proceedings of Historical Linguistics of the Caucasus* (Paris, 12–14 April, 2017) (pp. 168-177).

Tittel, S., Bermúdez-Sabel, H., and Chiarcos, C. (2018). Using RDFa to link text and dictionary data for Medieval French. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan (pp. 7-12).

Weingart, A., and Giovannetti, E. (2016). Extending the lemon model for a dictionary of old Occitan medico-botanical terminology. In *Proceedings of the European Semantic Web Conference (ESWC-2016)* (pp. 408-421). Springer, Cham.

Building ETCSANS: The Electronic Text Corpus of Syntactically Annotated Neo-Sumerian

Chiarcos, Christian

christian.chiarcos@gmail.com
Applied Computational Linguistics, Goethe Universität Frankfurt, Germany

Page-Perron, Emilie

emilie.page-perron@wolfson.ox.ac.uk
Wolfson College, University of Oxford, UK

Sumerian is a language of utmost importance to the cultural heritage of all humankind. As the very first written language, it has a vast literature spanning thousands of years (4th to 1st millenium BC) and ranges from genres as diverse as poems and songs over mythological and historical treatises, laws and letters to contracts and administrative records.

So far, much of this data remained underexplored and accessible to experts only. Although portals such as the Cuneiform Digital Library Initiative (CDLI) provided much of the textual data in digital form, very little of it had been translated and/or annotated, and if so, only at the level of morphological glosses (Cunningham et al., 2006; Zólyomi et al., 2008). In order to improve access to these texts for both a larger audience and machines, we developed an innovative annotation workflow (Chiarcos et al. 2018) and now provide the first syntactically annotated corpus of Sumerian, *ETCSANS: The Electronic Text Corpus of Syntactically Annotated Neo-Sumerian*. Our talk describes its creation, access strategies and usage scenarios. With respect to the syntactic annotation of cuneiform languages, some progress has recently been reported for Akkadian (Gordin et al. 2020, Sahala 2021), with similar challenges in orthography and sparsity, but for Sumerian, the situation is more complicated because of a number of unusual features of the language as a linguistic isolate (Bansal et al., 2021).

The ETCSANS corpus is the result of the project *Machine Translation and Automated Annotation of Cuneiform Languages (MTAAC, 2017-2020)*, and serves as a tool for the study of economy and society of the Neo-Sumerian period (2100-2000 BCE). ETCSANS complements two other corpora for Sumerian, the Electronic Text Corpus of Sumerian Literature (ETCSL, primarily Post-Sumerian), and the Electronic Text Corpus of Sumerian Royal Inscriptions (ETCSRI, all periods, for a limited domain), but these provide morphosyntactic annotations only. Prior to ETCSANS, no corpora have been published that also provide syntactic or semantic analyses for Sumerian. In order to facilitate the usability of ETCSANS for concrete research questions, we adopt a semantics-oriented approach to syntax, and to facilitate its usability beyond Assyriology, we followed the Universal Dependencies (UD). We discuss advantages and downsides of this approach and provide an overview and rationale for a number of notable cases where we had to deviate from UD conventions to account for philological and Assyriological requirements.

Aside from its Assyriological, philological and historical relevance, we addressed (and solved) a number

of typical challenges arising in DH projects, namely, the diversity and sparsity of existing data (dictionaries, morphological annotations), the limited capacities to create highly specialized annotations on a restricted budget, and the challenge to overcome technological barriers between different disciplines and infrastructures in the process. We systematically employ RDF as a middleware to integrate heterogeneous resources, automated pre-annotation, moderated crowdsourcing for quality control and the support for Linked Data as exchange and import mechanism. Machine learning also played a major role in the project, however this contribution focuses on the creation of data as a necessary prerequisite of machine learning. For this purpose, we had to resort to rule-based methods and manual refinements.

ETCSANS is accessible in different forms. We provide developer access for the annotated data via a public GitHub repository (https://github.com/cdli-gh). The native format is a tabular format (CDLI-CoNLL), from which Linked Data and TEI exports are being generated. The TEI export is designed for local querying by means of the TEITOK corpus management system (Janssen 2018). All three formats can be accessed from the CDLI API along with the associated inscribed artifacts, their original transliterations, object metadata, bibliography, and visualizations. These data are available in their native formats as well as in RDF, and linked with each other. All ETCSANS-related content, code, data and documentation is open source and available from our GitHub repositories. This also includes graphical user interfaces for search, visualization and manual annotation/revision of ETCSANS data as part of the larger CDLI framework.

With a total of 24,460 syntactically annotated texts, the ETCSANS core corpus covers about 22% of the overall Neo-Sumerian textual data. It consists of three subcorpora for which we could bootstrap syntax annotations in a semi-automated fashion, based on the domain (transaction subcorpus: 22,276 texts, 1,742,634 tokens, see Fig. 1 for an example), the possibility of annotation projection (parallel subcorpus: 1,572 texts, 46,321 tokens) or the existence of specialized morphological annotations (royal subcorpus: 612 texts, 9,133 tokens). Given the complexities of Sumerian writing and the specific nature of the Sumerian language, manual annotation focused on morphology, whereas manual syntax annotation is limited to a small evaluation corpus (11,220 tokens). The extended ETCSANS corpus contains another 47,476 texts (1,775,582 tokens), for which we provide automated annotations for morphology and named entities and a generic, rule-based annotator that exploits the explicit morphological marking of phrase structure boundaries in Sumerian morphology.

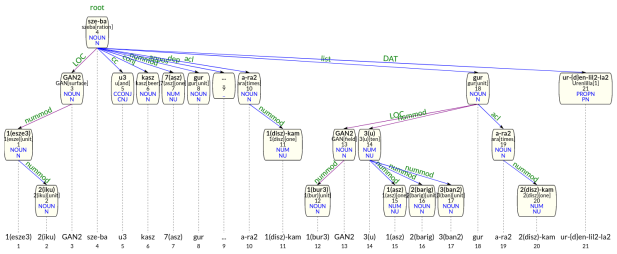


Fig. 1

Excerpt from a transaction text: "For 8 gan2 (~2.8 ha) [of acre] 7 gur (~2,100 l) rations and beer, first time, and for 18 gan2 (~6.5 ha) [of acre] 31.5 gur (~9,450 l), second time, for Ur-Enlilla." (P101040)

Overall, ETCSANS syntax annotation is largely derived from manual annotation or translations rather than directly manually created. Given the amount of data and the high degree of specialization required for doing the annotation, this is unavoidable, but from a methodological view, it presents a challenge that we would like to discuss with the wider DH audience, as we deal with (and delineate) different levels of trust and reliability, the possibilities (and limitations) for manual corrections and editorial decisions regarding the annotation, and, finally, the development of communication strategies to clarify potential and pitfalls of semiautomated annotations and their usage for future research in the humanities in a way that is transparent to scholars and laymen alike. At the time of writing, evaluation is still on-going. Preliminary results indicate that the different strategies we employed for bootstrapping annotations (two different approaches on rule-based annotation over manual morphology, annotation projection, ensemble combination of several such methods) differ greatly in the types of errors they produce, so that the focus of our current work is on developing methods to provide context-sensitive confidence scores for our annotations.

Bibliography

Bansal, R., et al. (2021). How Low is Too Low? A Computational Perspective on Extremely Low-Resource Languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, 44–59.

Chiarcos, C., et al. (2018). Annotating a low-resource language with LLOD technology. *Information*, 9:11, 290.

Cunningham, G., et al. (2006). *The Electronic Text Corpus of Sumerian Literature.* Oxford Text Archive Core Collection.

Gordin, S. et al. (2020). Reading Akkadian cuneiform using natural language processing. *PloS one*, **15:**10, e0240511.

Janssen, M. (2018). Dependency Graphs and TEITOK: Exploiting Dependency Parsing. In *International Conference on Computational Processing of the Portuguese Language,* Springer, Cham , pp. 470-478.

Sahala, A. (2021). *Contributions to Computational Assyriology*. PhD Dissertation, University of Helsinki.

Zólyomi, G., et al. (2008). *The Electronic Text Corpus of Sumerian Royal Inscriptions*. URL: http://oracc.museum.upenn.edu/etcsri/, Budapest (accessed 21 April 2022).

Computational approaches to literary periodization: an experiment in Italian narrative of 19th and 20th century

Ciotti, Fabio
fabio.ciotti@uniroma2.it
Università di Roma Tor Vergata, Italy

Theoretical background

Periodization is one of the fundamental topics of literary studies. As Rene Wellek puts it in one of the most notorious and important books of the last century theory of literature: "the concept of period is certainly one of the main instruments of historical knowledge", meaning, of course, literary-historical knowledge. (Wellek, 1956: 268). And yet is one of the most controversial and debated:

> It is virtually impossible to divide periods according to dates for, as [Jurij] Lotman points out, human culture is a dynamic system. Attempts to locate stages of cultural development within strict temporal boundaries contradict that dynamism. (Bassnett, 2013: 41)

How it comes that we can hypostatize the dynamic nature of cultural systems, superimposing on them a scalar chronology? How can we say, as Jameson puts it, that Ulysses is something that happened in 1922 (Jameson, 1971: 313)? Following Meneghelli (Meneghelli, 2013), we can individuate at least 4 critical issues, or even aporias in literary periodization:

1. Historical categories are related to cultural and social phenomena and overlap and interact in complex ways, determining *anisochronies* and *dischronies*;
2. Most if not all literary-historical categories have an ontological and trans-historical status and meaning embedded in them (take for instance Romanticism or Modernism);
3. Historical categories interact with and are dependent on geospatial ones, resulting in a multiplicity of asynchronous periodizations;
4. The notion of a historical category in literature is strictly associated with the canonical corpus of texts that are considered representative of a period, and the "dialectics between these two poles, period and canon, are complex and manifold" (Meneghelli, 2013: 3).

This last point is particularly relevant: literary periodization is usually the product of a process of generalization and synthesis, within a historical and social horizon, of the small-scale critical and interpretive practices that characterize the study of literary texts. Being bound to idiosyncratic hermeneutical practices and to the "epistemology of close reading", periodization suffers from all the pitfalls and limitations of that approach. In the last two decades, the landscape of literary and cultural studies has been enriched by a methodological perspective that is based on a quantitative approach. Among the various disciplinary labels that identify this current in studies, the most common is distant reading, introduced by Franco Moretti in his work on World Literature (Moretti, 2000) and subsequently extended to denote (even retroactively) the entire tradition of quantitative literary studies (Moretti, 2013; Underwood, 2019; Piper, 2018; Jockers, 2013). In this framework literary texts are elements of a population whose synchronic and diachronic characteristics should be empirically investigated on a molar scale, adopting statistical-probabilistic and computational methods. This would require the move from *interpretation* to *explanation* as the primary methodology of scholarly inquiry in the cultural and literary domains (Ciotti, 2021).

The possible contribution of a distant reading approach to the literary periodization problem, has been previously explored by some important studies, like (Jockers, 2013: chap. 6), (Piper, 2018: chap. 4), (Underwood, 2019) for English narrative fiction, all based on a supervised classification approach; while (Jannidis and Lauer, 2014) for German literature adopted a stylometric method. This paper explores the results of a mainly explorative and unsupervised analysis of a corpus of 19 th and 20th century Italian narrative fiction.

The corpus and the methodology

The main research hypothesis is if adopting computational quantitative methods, it's possible to identify time-based groupings in a set of texts distributed over a long historical period. Related to this there is the question of whether those eventual groupings are aligned with traditional historical periodization. The corpus consists of 660 Italian novels and short novels written between 1810 and 2000 of different aesthetic "levels", canonization status, genre, dimension, and authors' gender. Admittedly, this corpus is still far from being adequate and well balanced, mainly due to the uneven chronological distribution, but it is sufficient for an exploratory inquiry.

I wanted to test if the corpus can be clustered in a chronological sensible way using an algorithmic approach without presuming any prior categorization: this explains the preference for unsupervised methods in this research. In particular, I have adopted two different approaches, based on different assumptions, analytical techniques and features selection:

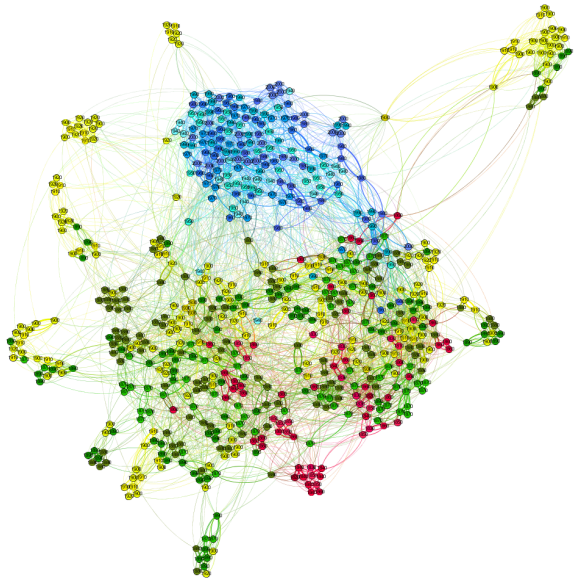
1. bootstrap consensus network, following (Eder, 2017) that applies phylogenetic consensus networks method to MFW based clustering;
2. lexicon-based text analysis and subsequent K-Means clustering of the results.

Fort the first experiment I have adopted the popular stylometric R package *Stylo* (Eder et al., 2016) to generate the consensus network dataset, conflating ten hierarchical clusterings generated comparing from 100 to 1000 most frequent words. The resulting output dataset is imported into the network analysis tool Gephi (Bastian et al., 2009) and each node is associated with an attribute that specifies its decade of composition. Then modularity is calculated with a resolution set at 4, resulting in 10 modules, and the final layout is generated applying the layout algorithm Force Atlas 2.

The second approach is based on the text analysis of the corpus with the tool *Linguistic Inquiry and Word Count* (*LIWC*) (Pennebaker et al., 2015). For each text, LIWC produces a vector containing the relative frequencies of various words-classes (E.g.: "Affect Words", "Cognitive Processes", "Perceptual Processes", "Biological Processes"....). The final output is a low dimensional document matrix, to which I have applied a K-Means clustering process, adopting the Python implementation of the algorithm in the *SciKit-Learn* library (Pedregosa et al., 2011). The choice of the number of clusters has been done evaluating 20 different models with *Elbow method* and *Silhouette Score* tests, which both indicated an optimal value of 4 clusters.

Results and future directions

The results of the stylometric approach represented in Fig.1 shows that the texts clustered in time sensible way are only those written in the second half of the 20 th century (that in the network have a blue color tone), while texts written in the first and second half of the 19 th century e in the first half of 20 th are not clearly separated, with the exception of the island in the upper right part of the graph, which is mostly composed of canonical Italian modernist texts.



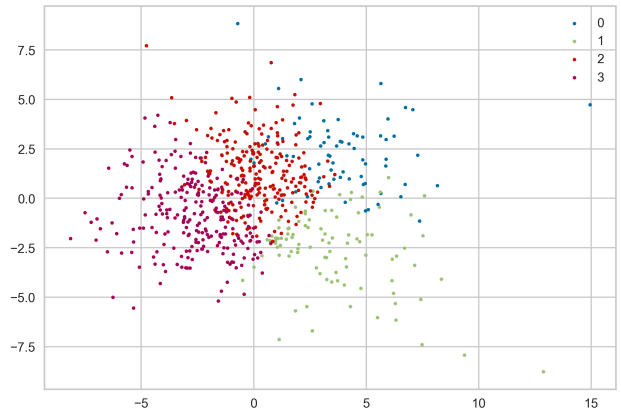
Consensus Network of the corpus: red 1810-1860; green 1860-1900; yellow 1900-1940; blue 1940-2010

This overall result is confirmed by the K-Means approach. For my analysis I have produced two matrices using the LIWC dictionary for Italian (Agosti and Rellini, 2007):

- a matrix with all LIWC lexical categories;
- a matrix limited to the following word categories: Verbs, Pronouns, orthographic signs of represented speech, and categories related to emotional and cognitive activity.

In this way I can test an additional hypothesis very common in literary-historical and critical scholarship: the linguistic sphere of the cognitive and emotional dimension is a characteristic feature of the evolution of narrative

along the nineteenth and twentieth centuries, namely it's a characteristic of the transition to the Modernism. While the K-Means clustering of the first matrix has very little relation to chronology, the second one provides clear indicators of temporal segmentation (Figure 2). Therefore, we can say that the incidence of the lexicon related to the sphere of thought/consciousness/emotivity is a signal of an evolutional pattern in the Italian novel. Anyway, also in this case the more clearly time-based cluster is that of the text written in the second half of the 20 th century.



K-Means clusters in the document matrix restricted to the cognitive/emotional lexicon

In conclusion, the first analysis in general confirms for Italian literature the limited role of the time dimension for clustering texts on a purely stylometric base already observed by (Jockers, 2013: chap. 6), as compared to authorship and even author gender. Instead, there is some evidence that quantitative empirical analysis partially confirms the relevance of cognitive/emotional lexicon in the evolution of literature. In this direction, I think that a fruitful development of the research will require a more effective way to identify the presence of cognitive/emotional attitudes in texts. To this end, we are training a streamlined Italian BERT language model to identify the relevant textual blocks, and the provisional results are promising.

Bibliography

Agosti, A. and Rellini, A. (2007). The Italian LIWC dictionary. *Austin, TX: LIWC. Net*.

Bassnett, S. (2013). *Translation Studies*. London: Routledge.

Bastian, M., Heymann, S. and Jacomy, M. (2009). Gephi - The Open Graph Viz Platform https://gephi.org/ (accessed 26 November 2018).

Ciotti, F. (2021). Distant reading in literary studies: a methodology in quest of theory. *TESTO & SENSO*(23).

Eder, M. (2017). Visualization in stylometry: Cluster analysis using networks. *Digital Scholarship in the Humanities*, **32**(1): 50–64 doi:10.1093/llc/fqv061.

Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R: A Package for Computational Text Analysis. *The R Journal*, **8**(1): 107–21.

Jameson, Fredric. (1971). *Marxism and Form : Twentieth-Century Dialectical Theories of Literature*. Princeton (N.J.): Princeton University Press.

Jannidis, F. and Lauer, G. (2014). Burrows's Delta and Its Use in German Literary History. In Erlin, M. and Tatlock, L. (eds), *Distant Readings. Topologies of German Culture in the Long Nineteenth Century*. Rochester: Camden House, pp. 29–54 gerhardlauer.de/index.php/download_file/view/335/1/.

Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History*. (Topics in the Digital Humanities). University of Illinois Press

helli, D. (2013). Periodization, Comparative Literature, and Italian Modernism. *CLCWeb: Comparative Literature and Culture*, **15**(7) doi:10.7771/1481-4374.2386. https://docs.lib.purdue.edu/clcweb/vol15/iss7/12 (accessed 8 December 2021).

Moretti, F. (2000). Conjectures on World Literature. *The New Left Review* http://newleftreview.org/A2094.

Moretti, F. (2013). *Distant Reading*. London: Verso http://www.amazon.de/Distant-Reading-Franco-Moretti/dp/1781680841.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**: 2825–30.

Pennebaker, J. W., Boyd, R. L., Jordan, K. and Blackburn, K. (2015). The Development and Psychometric Properties of LIWC2015. University of Texas at Austin doi:10.15781/T29G6Z. http://hdl.handle.net/2152/31333 (accessed 8 December 2021).

Piper, A. (2018). *Enumerations: Data and Literary Study*. Chicago ; London: The University of Chicago Press.

Underwood, T. (2019). *Distant Horizons: Digital Evidence and Literary Change*. Chicago: The University of Chicago Press.

Wellek, R., Warren, Austin,, (1956). *Theory of Literature*. New York: Harcourt, Brace & World.

Building a journal for the digital era: the Journal of Digital History

Clavert, Frédéric
frederic.clavert@uni.lu

Centre for Contemporary and Digital History, University of Luxembourg, Luxembourg

Fickers, Andreas

andreas.fickers@uni.lu
Centre for Contemporary and Digital History, University of Luxembourg, Luxembourg

The Journal of Digital History (JDH) was launched in October 2021. It is a joint effort of the Centre for Contemporary and Digital History (University of Luxembourg) and De Gruyter Publishing. For two years, the building of the JDH required the collaboration of a fully interdisciplinary team. The JDH is based on the concept of multilayered articles, that was defined taking into consideration a full genealogy of rethinking academic publications in the digital era.

In 1999, Robert Darnton (Darnton, 1999) explained his vision of the post-web scholarship, through the multilayered book, that would make sense of the emergence of the Internet, of possible new formats of digital storytelling and narration. He described this vision of a book as layers 'arranged like a pyramid', a pyramid which would bring together exchanges with readers, pedagogic contributions, historiographical discussions, primary sources, extended version of the content and, on top of that, the book itself.

Darnton had not been the only one to think about academic writing in the digital era. In history, *The Valley of the Shadow* is an early attempt at finding new forms of critical engagements with historical scholarship. Both directors of this project insisted on the possibilities of 'mature hypertextual history' (Ayers, 2001) that enables readers to 'follow the logic of our thinking' (Thomas, 2004).

With the JDH, we are enabling authors to go beyond those first attempts and enabling readers to not only read the article itself but understand the research practices and multiple decisions that finally led to the writing of the article. Producing here transparency about how the digital interferes in the iterative research process is a matter of epistemological imperative. This is linked with digital hermeneutics, a combination of critical digital skills with a self-reflexive approach (Fickers, 2020).

The multilayered article is the JDH's answer to this challenge of making explicit how the production of historical knowledge by means of digital tools and technologies is the result of co-construction of the 'epistemic object' of historical investigation through human-machine interaction (Rheinberger, 2010; Spencer, 2019). We defined three layers: narrative – allowing authors to make use of all the web's possibilities in terms of transmedia storytelling – , hermeneutics – methods, tools and code – , and data – the dataset.

To implement the multilayered article, the JDH team reviewed several existing open source software and finally chose code notebooks. Though with some short ends – poor structure, no proper way to cite literature – , notebooks offered many possibilities: markdown (that allows lightweight structuration); support for the main computing languages used in digital humanities; interaction through the use of myBinder. Jupyter notebooks became the central piece of the JDH infrastructure (figure 1). Datasets are stored on a dataverse instance. Notebooks are stored on a git server – github at the moment. We then set up a conversion pipe from the notebooks to the journal's website.

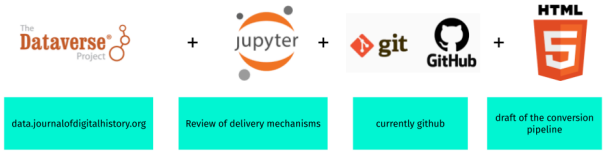


Figure 1
The JDH infrastructure

The reader will experiment with multilayered articles thanks to an interface that crosses the different layers in different ways. When arriving on the landing page of the article, the reader will first see the narrative layer (figure 2) and can open the hermeneutic cells to punctually see the methods or the code written (figure 3). Another possibility is to choose the 'hermeneutics first' layout (figure 4). The last option is to launch the notebook in the myBinder service from the submenu 'Data' (figure 5), where readers may interact with the code (and hence with the dataset) which is also a way to question the article's hypotheses.

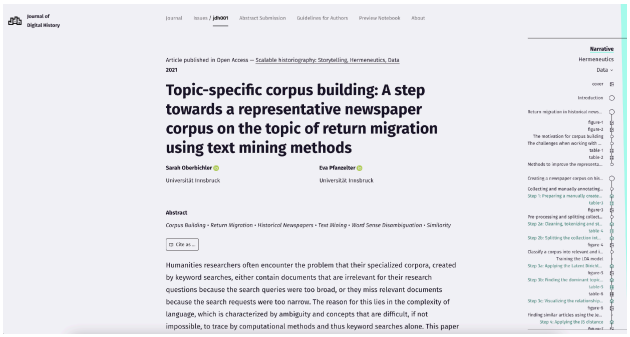

Topic-specific corpus building: A step towards a representative newspaper corpus on the topic of return migration using text mining methods

Figure 2
The narrative layer

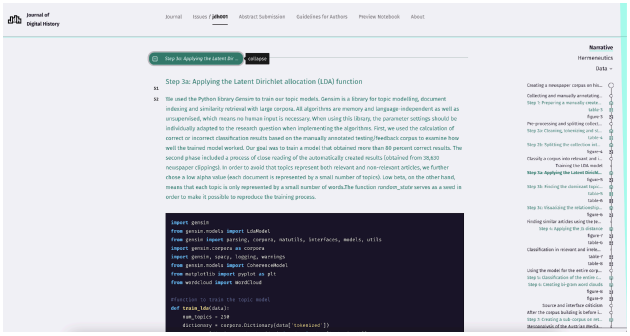


Figure 3
Hermeneutics cells through the narrative layer

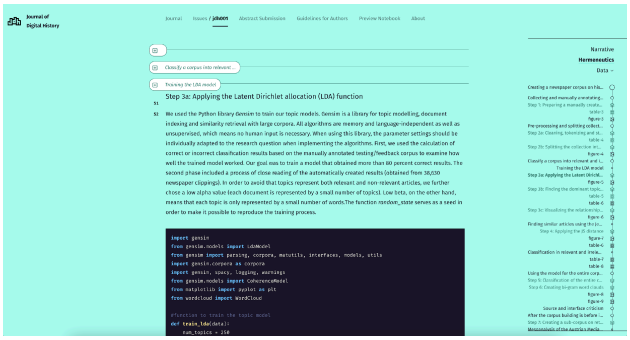


Figure 4
The hermeneutics layer

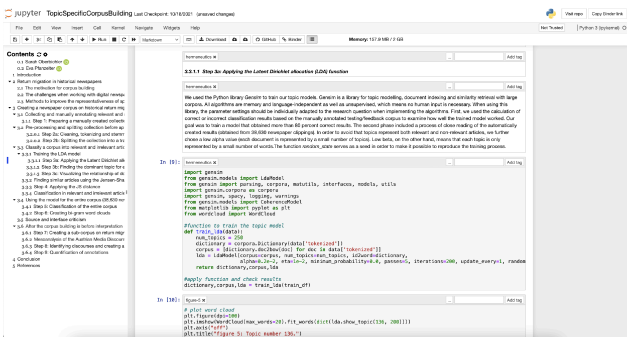


Figure 5
Interacting with the article through myBinder

All along the process of designing and developing the infrastructure, we entered a process of co-building the journal's infrastructure with authors through the organisation of workshops. The idea to use notebooks, for instance, has been discussed with authors during one of those. When abstracts are submitted, individual meetings with authors allowed us to adapt to their needs.

The interdisciplinarity of the JDH's team at the C2DH and its horizontal functioning have also been a strong strength: the cumulative competences of designers (2), developers (2), server admin (1), copy editors (1) and historians (2) allowed us to build a publishing infrastructure within 18 months, from the use of open source bricks.

We are currently facing several limitations:

The JDH is open access and favours open data, but legal restrictions, especially for the 20th and the 21st century, restrict this policy as well as technical limitations when the notebook deals with very large amounts of data;

- Finding reviewers who are literate in history and in computing is not easy;
- It is still unclear how code and datasets should be evaluated;
- Citations are dealt with by zotero, through a notebook extension (cite2c) that deserves much more development.

The JDH's next perspectives are editorial: we will experiment with the notion of 'issue'. A JDH issue is a stream of articles. We do not publish all articles of an issue at the same time, because the JDH team's workload for each article is high. We would now like to investigate open-ended issues, on teaching digital history for instance.

Bibliography

Ayers, E. (2001). The Pasts and Futures of Digital History. *History News*, 56(4): 5-9.

Darnton, R. (1999). The New Age of the Book. *New York Times*, March.

Fickers, A. (2020). *Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur digitalen Forensik?* ZZF Centre for Contemporary History: Zeithistorische Forschungen, https://doi.org/10.14765/ZZF.DOK-1765 (accessed 22 March 2022).

Rheinberger, H-J. (2010). *On Historicizing Epistemology: An Essay*. Stanford University Press.

Spencer, M. (2019). The Difference a Method Makes: Methods as Epistemic Objects in Computational Science. *Distinktion: Journal of Social Theory*, 20(3): 313–27, https://doi.org/10.1080/1600910X.2019.1610018 (accessed 22 March 2022).

Thomas, W. G. II. (2004). Computing and the Historical Imagination. In Susan, S. *et al.*, *Companion to Digital Humanities*, Oxford: Blackwell, 56-68.

New ways to read Greek and Persian epic and to explore diverse cultures

Crane, Gregory

gregory.crane@tufts.edu
Tufts University, United States of America

Shamsian, Farnoosh
shamsian@informatik.uni-leipzig.de
Leipzig University, Germany

Babeu, Alison
alison.babeu@tufts.edu
Tufts University, United States of America

Tauber, James
jtauber@jtauber.com
Eldarion.com, United States

Wegner, Jacob
jwegner@eldarion.com
Eldarion.com, United States

Our work explores the hypothesis that a new mode of reading is taking shape, one in which dense, machine actionable annotations allow readers to work directly and effectively with sources in languages that they do not know – a new middle space between reliance on translation and mastery of the source text (Crane et al. 2019, Crane 2019). This hypothesis has substantial potential importance for our ability to use source texts to explore cultural diversity in general and the diversity of Asian cultures in particular. Our particular work focuses on two challenges for a traditionally Eurocentric subject, Classics (or Classical Studies), which is still used to describe the study of Greco-Roman culture. On the one hand, university students without training in Greek and Latin in secondary school have difficulty mastering the languages and learning about the subject. In spring 2021, the Princeton Classics Department provoked controversy when it made it possible for majors to study Greco-Roman antiquity without learning any Greek or Latin – too few students, especially students of color, had access to Latin, much less Greek, before college (Wood 2021). At the same time, Classics and Classical Studies are far too narrow – we must include other classical languages – Sanskrit, Classical Chinese, Classical Arabic, etc. – if we are to continue using these terms. We report on work that addresses both challenges.

In order to explore this broad topic, we chose to focus on two complementary corpora in two Classical languages: the Homeric epics in Ancient Greek and the *Shahnameh* in early-modern Persian (Firdawsī 1430, 1988). The goal is both to support Persian speakers who wish to work directly with Homeric Epic and English speakers who wish to engage directly with the *Shahnameh.* In the case of Persian culture, the links with Greco-Roman culture are deep, the information in Greek sources about ancient Persian history is extensive, and the influence of Greek philosophy, medicine and science are extensive. At the same time, few institutions in Europe and North America, for example, teach modern Persian, much less the early modern Persian of the Shahnameh. We hope to increase the role that the Persian epic, in Persian as well as in translation, plays beyond the Persian speaking world.

The use of dense linguistic annotation to make sources accessible to a broader audience is, of course, hardly new. In his late seventeenth-century description of Ottoman Turkish, Arabic, and Persian, for example, Franciszek Meniński (1680) introduced Persian poetry to a European audience by transliterating a passage from a poem by Hafez, providing a word-by-word translation, and providing detailed explanations of the metrical, morphological, and syntactic function of each word. Contemporary linguists depend upon exhaustively annotated text to work across sources from the thousands of languages, ancient as well as modern, in the human record (Werning 2009).

Digital methods, however, fundamentally change our ability (Berti 2019, Schulz 2021). First, we can use natural language processing pipelines such as Stanza and Spacy (Papantoniou and Tzitzikas 2020), multilingual language models such as BERT (Bamman and Burns 2020), machine translation (Kontogianni et al. 2020 *)*, most effective for now between modern languages (Bowker 2021), and similar openly licensed resources. In the work that we present, we document the categories which we have found as starting points to augment machine readable texts. The Homeric epics have provided a useful starting point because a particularly rich set of preexisting, open digital resources are available upon which to build. The work with Homeric epic provides us with the framework upon which we are building work with the Shahnameh. We will report on the work with Homeric Greek and summarize progress with pre-modern Persian (building on Pizzi 1881).

Exhaustive Metrical Analysis and accompanying sound: Machine-actionable metrical analyses (Schoisswohl and Papakitsos 2020) for every syllable in every line of the *Iliad* and the *Odyssey* and readings for a substantial portion of the epics are available under an open license (Chamberlain 2021). From the time they learn the Greek alphabet learners engage immediately with the Homeric epic as performed poetry, following metrical diagram and performance.

Treebanks document features such as the dictionary entry, part of speech, and syntactic function of every word in a source (Keersmakers, 2019). Treebanks are available for both the *Iliad* and *Odyssey* and for more than 1 million words of Greek and Latin (Bamman and Crane, 2011; Celano 2019). We can use these to identify and quantify

grammatical structures that students will encounter in the corpus that they will learn.

Paradigm information identifies the morphemes within each individual word and allows learners to see which morphological patterns are most common and to prioritize their learning.

Born-digital aligned translations are created from the start to expose the linguistic structures of a source. From the first lessons, learners can explore the meaning of vocabulary by seeing passages where these words appear (Palladino 2020, Palladino et al. 2021). They can focus on words introduced in a lesson but, in using the aligned translation, they gain constant incidental exposure to words that they have not learned. With translation, the problem of the learner language becomes far more pressing and we can report on the varying challenges of articulating Greek language with translations into English and Persian (Foradi 2020).

Grammatical explanations explain the patterns that learners encounter in the annotations (Mugelli 2021) described above (e.g., the various uses of the dative in Greek or the way we translate the imperfect tense). Grammatical explanations build upon the aligned translations. Grammatical explanations cannot simply be translated but must be adapted to bridge the gap between the target language and the learner language. We report on the differences that arise for speakers of Persian and English.

We will report upon experiences of learners from both Iran and the United States, upon our experiences opening historical students to new audiences (e.g., Classics majors who do not know Greek or Latin) and cross-cultural explanation of content (e.g., Homeric Epic for Persian speakers and Persian poetry such as the Shahnameh for English speakers).

Bibliography

Bamman, D., Burns, P. (2020). “Latin BERT: A Contextual Language Model for Classical Philology.” https://arxiv.org/abs/2009.10053

Bamman, D., Crane, G. (2011). “The Ancient Greek and Latin Dependency Treebanks,” in Language Technology for Cultural Heritage, ed. Caroline Sporleder, Antal van den Bosch, and Kalliopi Zervanou, Theory and Applications of Natural Language Processing (2011), pp. 79–98.

Berti, Monica, ed. (2019). Classical Philology. Ancient Greek and Latin in the Digital Revolution. Berlin: De Gruyter Saur.

Bowker, Lynne. (2021). "Digital humanities and translation studies." Handbook of Translation Studies: Volume 5 5 (2021): 37-.

Celano, Giuseppe. (2019). “The Dependency Treebanks for Ancient Greek and Latin”. In Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution, edited by Monica Berti, 279–298. Berlin: De Gruyter Saur.

Chamberlain, David. (2021). “A Reading of Homer (work in progress).” Greek and Roman Verse, https://hypotactic.com/my-reading-of-homer-work-in-progress/. Accessed 8 December 2021.

Crane, G. R; Shamsian, F.; et al . (2019). “Confronting Complexity of Babel in a Global and Digital Age.” DH2019: Digital Humanities Conference, Book of Abstracts (2019), pp. 127–138.
https://dev.clariah.nl/files/dh2019/boa/0611.html

Crane, Gregory. (2019). “Beyond Translation: Language Hacking and Philology.” Harvard Data Science Review 1, no. 2. https://doi.org/10.1162/99608f92.282ad764.

Firdawsī, A. & Khaleghi-Motlagh, D. (1988). The Shahnameh: Book of kings. New York: Bibliotheca Persica.

Firdawsī, A. & Ja'Far, P. C. (1430) The Book of Kings. Tehran: Cultural Heritage Organization. [Pdf] Retrieved from the Library of Congress, https://www.loc.gov/item/2021667287/.

Foradi, Maryam. (2020). Engagement with Classical Literature in the Framework of a Citizen Science Project Using Translation Alignment: Date Accuracy and Pedagogical Effectiveness, [Doctoral dissertation, University of Leipzig]

Keersmaekers, Alek. (2019). “Creating, Enriching, and Valorizing Treebanks of Ancient Greek.” 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019),
https://doi.org/10.18653/v1/W19-7812

Kontogianni, A., et al. (2020). “Computer-Assisted Translation of Egyptian-Coptic into Greek.” Journal of Integrated Information Management, http://ejournals.uniwa.gr/index.php/JIIM/article/view/4470

Meniński, F. (1680). Thesaurus Linguarum Orientalium Turcicae, Arabicae, Persicae, Praecipuas earum opes a Turcis peculariter usurpatas continens, Vienna: Franciscus a Mesgnien Meninski.

Mugelli, Gloria et al. (2021). “Learning Greek and Latin Through Digital Annotation: The EuporiaEDU System.” In: Teaching Classics in the Digital Age. Kiel: Universitätsverlag Kiel | Kiel University Publishing. S. 25–36. (= Think! Historically: Teaching History and the Dialogue of Disciplines). Online unter: https://macau.uni-kiel.de/receive/macau_mods_00001367.

Palladino, Chiara (2020). ”Reading Texts in Digital Environments: Applications of Translation Alignment for Classical Language Learning.” The Journal of Interactive Technology Pedagogy, Issue 18, December 10, 2020, https://jitp.commons.gc.cuny.edu/reading-texts-in-digital-environments-applications-of-translation-alignment-for-classical-language-learning/

Palladino, C., Foradi, M., Yousef, T. (2021). "Translation Alignment for Historical Language Learning: A Case Study." Digital Humanities Quarterly, 15 (3), http://digitalhumanities.org/dhq/vol/15/3/000563/000563.html

Papantoniou, K., Tzitzikas, Y. (2020). "NLP for the Greek Language: a Brief Survey." SETN 2020: 11th Hellenic Conference on Artificial Intelligence, pp. 101-109. https://doi.org/10.1145/3411408.3411410

Pizzi, I. Manuale della lingua persiana. Grammatica, antologia e vocabolario, Leipzig, W.

Gerhard (1881). Available at https://archive.org/details/manualedellalin00pizzgoog/

Schoisswohl, O., Papakitsos , E. C. (2020). "Automated metric profiling and comparison of Ancient Greek verse epics in Hexameter." Linguistik Online, 103(3), 159–177. https://doi.org/10.13092/lo.103.719

Schulz, Konstantin (2021). "Natural Language Processing for Teaching Ancient Languages." In: Teaching Classics in the Digital Age. Kiel: Universitätsverlag Kiel | Kiel University Publishing. S. 37–48. (= Think! Historically: Teaching History and the Dialogue of Disciplines). Online unter: https://macau.uni-kiel.de/receive/macau_mods_00001368.

Werning, Daniel A. (2009) "Glossing Ancient Egyptian. Suggestions for Adapting the Leipzig Glossing Rules." Lingua Aegyptia. Journal of Egyptian Language Studies, https://www.academia.edu/1484975/Glossing_Ancient_Egyptian_Suggestions_for_Adapting_the_Leipzig_Glossing_Rules.

Wood, G. (2021, June 9). Princeton Cancels Latin and Greek. The Atlantic. Retrieved December 10, 2021, from https://www.theatlantic.com/ideas/archive/2021/06/princeton-greek-latin-requirement/619136/

Good Grief! Encoding, Quantifying, and Analyzing *Peanuts* in the Classroom

Croxall, Brian
brian.croxall@byu.edu
Brigham Young University, United States of America

At first, the *Peanuts* comic of 23 January 1952 seems easy to understand. Charlie Brown and Schroeder, two of Charles M. Schulz's most familiar characters, talk about the music the latter is playing. Upon looking closer, the visual representation is far more complex than what is expected: Schulz reproduces the opening stanzas of Beethoven's "Grosse Sonate für das Hammerklavier" (Piano Sonata No.29, Op.106); Schroeder, responding to Charlie Brown's question about the piece, replies in German; and his speech is rendered in *Fraktur*; Charlie Brown's response is a single typographic character; and Schulz signs his name in the final panel in *Fraktur*.

How do you train students to see the complications in this strip? One solution would be to have them read Scott McCloud's accessible classic *Understanding Comics*, Hilary Chute's more recent *Why Comics?*, or the audacious *How to Read* Nancy by Paul Karasik and Mark Newgarden. But another solution would be to have them read the comics very closely through the act of encoding them and to then quantify their findings. In this presentation, I will discuss recent courses I have taught that have taken a digital humanities approach to Schulz's corpus.

Peanuts first appeared in October 1950. As a "space-saving strip," which meant editors could arrange its panels in different ways on the comics page, Schulz's strip may have appeared destined to be quickly forgotten (Schulz, "An Interview" 316). Instead, he continued drawing strip after strip for the next 50 years. When the final Peanuts strip was published on 13 February 2000 (the morning after Schulz had passed away), he had singlehandedly drawn, inked, and lettered 17,897 strips, resulting in what Robert Thompson called the "longest story ever told by one human being" (qtd. in Boxer). *Peanuts* influenced not only other comics but culture as a whole, with Charlie Brown, Snoopy, and the others being some of the most recognizable characters on earth. Compared to the strips that preceded Schulz's, such as *The Yellow Kid* (1895-1898), *Krazy Kat* (1913-1944), and *Pogo* (1948-1975), or those that followed it, including *The Far Side* (1979-1995) and *Calvin and Hobbes* (1985-1995), Schulz's style of both illustration and humor was minimal. But as the preceding paragraph suggests, simplicity does not preclude complexity.

For the last two years, I have taught a semester-long "Research in Digital Humanities" course that focuses on Schulz's work. While we closely read more than four years worth of strips, we extend our understanding of Schulz's work through the application of different digital humanities methods. First, we have collectively begun a digital edition of *Peanuts*, encoding strips according to both TEI and the Comic Book Markup Language (CBML; see Walsh). Encoding strips allows us to precisely record features of strips, including setting, weather, activities the characters are engaged in, and whether the strips are tied to a particular holiday (see Croxall et al., *Peanuts Taxonomy*). We tag characters who are named in the strip as well as place names, and record speech, sounds, types of speech bubbles, and diegetic text that appears in the strip. Throughout this process, the choices of what we encode are collectively determined by the class (see Croxall et al, *Peanuts Encoding Editorial Decisions*). We discuss the encoding regularly in

class sessions, as precision requires us to look carefully at both the comic and the TEI and CBML guidelines.

While we slowly build this digital edition, we also use a data set of transcriptions (similar to *alt-text*) harvested from the *GoComics* website that gives us the ability to search across all 50 years of the comics' history. With these two different data sets—one small and precise, the other broad and varied—we then engage in distant reading approaches. The first year of the class was dedicated to determining character co-occurrence within the strips and analyzing the networks among Schulz's characters (see Fig. 1). The second year found us using stylometry to determine whether the different characters had distinct speaking patterns. In the upcoming semester (Winter 2022), we will consider the bounds of season for both weather and sports.

In this presentation, I will discuss both our methods of encoding and the results of our quantifications. While drawing on the work of others who have examined comics in the terms of digital humanities (see Whitson and Salter; Dunst et al.), I extend the conversation for the first time into the convergence of comics, DH, and pedagogy.

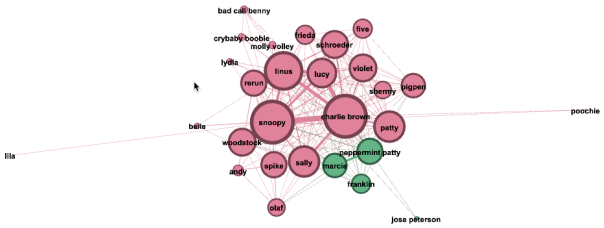


Fig. 1,
Network Analysis of In-strip Character Co-Occurrence

Bibliography

Boxer, S. (2000). Charles M. Schulz, 'Peanuts' Creator, Dies at 77. *New York Times*. 14 February 2000, https://www.nytimes.com/2000/02/14/arts/charles-m-schulz-peanuts-creator-dies-at-77.html?pagewanted=all&src=pm (accessed 27 April 2022).

Chute, H. (2017). *Why Comics? From Underground to Everywhere*. New York: Harper.

Croxall, B. et al. (2020-2022). *Peanuts Encoding Editorial Decisions*. Google Docs, https://docs.google.com/document/d/1-4u0SHY5PW16sfcEqL8mOYl-x2JB0piXowOOpLYGn7I/edit?usp=sharing (accessed 27 April 2022).

Croxall, B. et al. (2020-2022). *Peanuts Taxonomy*. GitHub, https://github.com/briancroxall/peanuts-taxonomy (accessed 27 April 2022).

Dunst, A. et al. (2018). *Empirical Comics Research: Digital, Multimodal, and Cognitive Methods*. New York: Routledge.

Karasik, P. and Newgarden, M. (2017). *How to Read* Nancy*: The Elements of Comics in Three Easy Panels*. Seattle: Fantagraphic Books.

McCloud, S. (1993) *Understanding Comics: The Invisible Art*. New York: William Morrow.

Schulz, C. (2004). "An Interview with Charles M. Schulz." *The Complete Peanuts: 1950-1952*. Seattle: Fantagraphic Books, pp. 304-337.

Schulz, C. (2014-2016). *The Complete Peanuts*. 25 vols. Seattle: Fantagraphic Books.

Schulz, C. 23 January 1952. *GoComics*, https://www.gocomics.com/peanuts/1952/01/23 (accessed 27 April 2022).

Walsh, J. A. (2012). "Comic Book Markup Language: An Introduction and Rationale." *Digital Humanities Quarterly* 6.1, http://www.digitalhumanities.org/dhq/vol/6/1/000117/000117.html (accessed 27 April 2022).

Whitson, R. T. and Salter, A. (2015). "Introduction: Comics and the Digital Humanities." *Digital Humanities Quarterly* 9.4, http://digitalhumanities.org/dhq/vol/9/4/000210/000210.html (accessed 27 April 2022).

Multilinguals Write Back: Modeling Language, Politics and Identity in Philippine Social Media

Cruz, Frances Antoinette
Frances.Cruz@uantwerpen.be
University of Antwerp/University of the Philippines Diliman

Kestemont, Mike
mike.kestemont@uantwerpen.be
University of Antwerp

Research problem statement

This paper documents the intersections of language and the public sphere through a cross-sectional model of comments on public Facebook (FB) pages of selected newspapers from 2015-2019. The analysis examines

language use in discussions of current events, differences between national and regional newspapers, and social media insights into the conduct of public discourse. A common observation in the Philippines is, while English is used for official documents, tertiary education, and national broadsheets, oral discussions tend to involve Filipino or regional languages (Gonzalez, 1998). Class and political differences are also heavily associated with language use and media preferences (Kusaka 2017). Social media, with its informal written language, yet socially and politically relevant content, may thus offer insights into contemporary language use and public engagement with media.

Methodology

Comments on public FB pages of selected national and regional newspapers (*Manila Bulletin*, *Manila Times*, *Philippine Daily Inquirer*, *Philippine Star*, *Cebu Daily News*, *Mindanews*, *Mindanao Times*, *Sun Star Cebu*, *Sun Star Davao*, and *The Freeman*) were captured through *Facepager* (Jünger & Keyling 2019). The corpus is taken from a separate project on Muslim identities in the Philippines and consists of data from selected months in 2015, 2017, and 2019.

While language identification is a common task in natural language processing, most of the available software casts this as a multi-class classification task, where only a single language can be assigned to a textual document. In the present case, involving code-switching as a signature feature, multiple languages can be simultaneously present in a document, thus turning this task effectively into a multilabel classification problem. We finetuned a bespoke multilabel classifier on top of a pretrained BERT (Devlin et al., 2019) feature extractor (the 'bert-base-multilingual-uncased' model). We only considered messages where the target language(s) could be identified and divided these into a train, validation and test set of 10,000 social media messages (containing 7,304, and 2 x 913 instances respectively). We manually annotated for the presence/absence of Tagalog/Filipino, Cebuano, and English respectively. We compared the performance of this SOTA approach to a simpler baseline, consisting of a conventional multitarget classifier in the form of a random forest (RF) (Pedregosa et al., 2011) of 16,156 manually annotated entries (with a train, validation and test set of 12,998 entries, and 2 x 2,294 instances), trained on top of a TF-IDF representation of a vocabulary character n-grams (for $2 \leq n \leq 6$) (see details Table 2 below). Finally, we applied the BERT language detector (with the weights that optimized the validation performance) to the unseen data.

	BERT Train, Validation and Test Set (10,000 entries)	RF Train, Validation and Test Set (16,156 entries)
Cebuano	1648	2345
Tagalog	3213	5584
English	2950	4788
Cebuano and Tagalog (Bislog)	50	69
Cebuano and English (Bislish)	238	370
Tagalog and English (Taglish)	15	1661
Cebuano, Tagalog and English	1016	20

Table 1
Number of entries per language in Test, Train and Validation Sets

	BERT	Random forest
Code-level accuracy	95.50	93.82
Message-level accuracy	90.36	85.76
Overall F1	93.97	91.54
Cebuano/Bisaya (Acc / F1)	94.30 / 93.85	92.44 / 91.55
Tagalog/Filipino (Acc / F1)	95.07 / 94.84	92.77 / 92.41
English (Acc / F1)	97.15 / 93.22	96.27 / 90.65

Table 2
Test Accuracies

Findings

Both classifiers were able to demonstrate similar general trends: Tagalog/Filipino entries were the most common overall (Figures 1 and 2). Regionally, more diverse language use is noticeable. Cebuano comments were prominent in at least two Cebu-based newspapers, while the Mindanao Times and Sun Star Davao featured Tagalog/Filipino as the most-used language, followed by English and Cebuano (Figures 3 and 4), suggesting the usage of Tagalog/Filipino even where Visayan languages are prominent. Despite the presence of monolingual English-language newspapers, the fact that current events are **written about** and responded to by a multilingual Philippine public sphere is often obscured. Social media thus offers opportunities not only for re-thinking monolingual norms in media, but may also act as a forum for revitalizing written forms of regional languages, while acting as a potent corpus and resource for codeswitching and informal written language for automatic language identifiers.

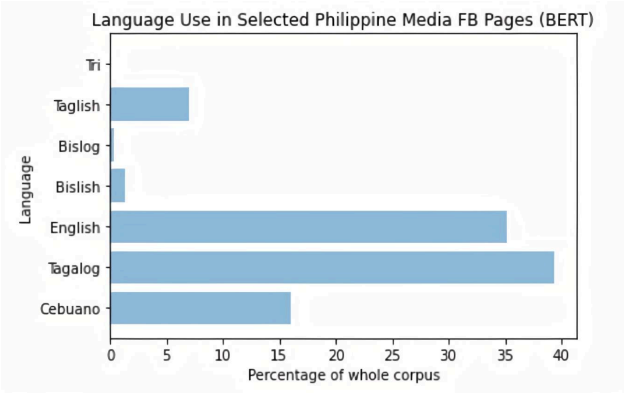


Fig. 1
Language Use (BERT)

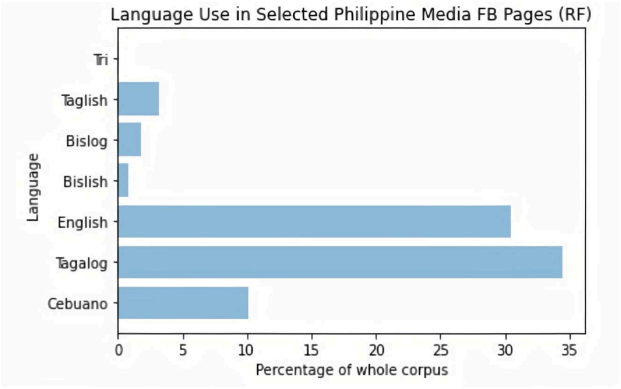


Fig. 2
Language Use (RF)

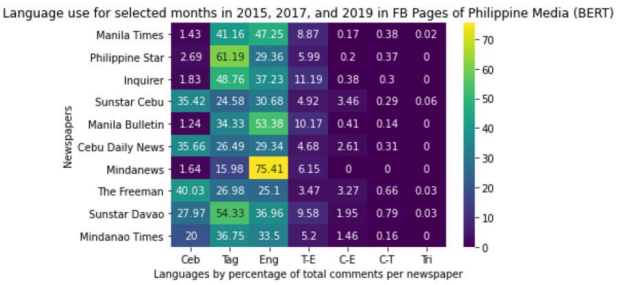


Fig. 3
Languages per Newspaper (BERT)

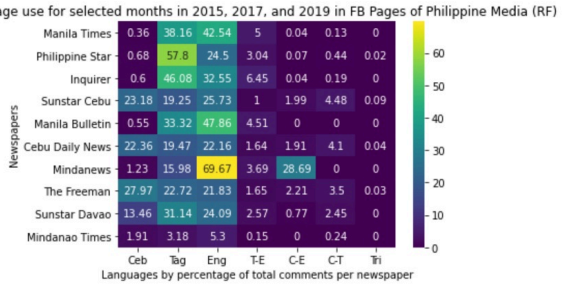


Fig. 4
Languages per Newspaper (RF)

Bibliography

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, Minneapolis, MN, June 2019, pp. 4171–4186.

Gonzalez, A. (1998). The Language Planning Situation in the Philippines. *Journal of Multilingual and Multicultural Development*, *19*(5 & 6): 487–525.

Jünger, J. and Keyling, T. (2019). *Facepager: An application for automated data retrieval on the web*. https://github.com/strohne/Facepager/ (accessed April 7 2022).

Kusaka, W. (2017). *Moral Politics in the Philippines*. Singapore/Japan: NUS Press & Kyoto University Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Bertrand, T., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., David Cournapeau, Brucher, M., Perro, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**: 2825–2830.

Strategies for DH awareness-raising activities: the challenges of DH Awards

Cummings, James
james.cummings@newcastle.ac.uk
Newcastle University, United Kingdom

Introduction

This long paper explores the annual Digital Humanities Awards (DH Awards), run as an openly-nominated and openly-voted DH awareness-raising activity at http://dhawards.org/. Specifically, it looks at possible solutions to criticisms and concerns by foregrounding a number

of problematic aspects of DH Awards in its history and future. Although many people praise these annual awards for raising awareness of DH outside the usual disciplinary bubble, there are significant criticisms of them, including being merely a popularity contest, horribly skewed to the western and anglo aspects of DH, more likely to highlight well-funded projects, and their over-estimated importance by those outside DH in tenure and promotion reviews.

This paper is presented by the founder and main instigator of DH Awards and seeks to build on discussions of these issues. This presentation benefits from a privileged position of having access to all of the anonymized nomination and voting data for the entire run of DH Awards (2012 to present), as well as informal feedback which it will use in an aggregated anonymized manner. This self-reflexive paper does not seek to dismiss criticisms, but to engage with them in good faith in order to understand DH Awards' role in the community, and how best to ameliorate any issues raised.

Background and History

The underlying conception for DH Awards was born during the Roberto Busa Prize lecture 'Living with Google: Perspectives on Humanities Computing and Digital Libraries' by Professor Susan Hockey at the ACH/ALLC 2004 conference in Gothenburg (Hockey, 2005). Professor Hockey was an excellent and deserving choice for that prize, which recognizes "outstanding lifetime achievements in the application of information and communications technologies to humanities research." [1] However, listening to the lecture prompted naive questions about the awarding of the Roberto Busa Prize: who made the decisions, who was allowed to nominate and, crucially, who else might have been nominated? While, in reality, it is likely only a small number of candidates were considered, the lack of transparency worried me and especially not hearing about others that may have been nominated. This sparked the idea for an openly nominated and openly voted grassroots approach to some DH awards which would act as a way to showcase DH resources. My ignorance of the process at that time meant that it felt far from the more transparent and representative protocols that are documented on the ADHO website today. [2] DH Awards creation was also based in discussions at the time of the economics of academic career structures and how community-based catalysts could help support this. [3] When this idea resurfaced in 2012, an international group was approached to act as a very light-touch nominations committee. The members of this group were specifically chosen in a laudable attempt to have a geographic and cultural spread but were also existing DH colleagues who were likely to agree to be on the committee. The creation of a website, nominations and voting forms, and the running of the nominations and voting followed swiftly without much reflection. The process evolves each year but the DH Awards follows a consistent, if haphazard, pattern each year. In autumn the committee is approached to see if they wish to continue and suggestions for modification of categories or other aspects, nominations open in December until late January, then the nominations committee checks them to see that they:

1. Are vaguely 'Digital Humanities' in the loosest sense
2. Had some update/change of any sort that year
3. Are nominated in the right category
4. Able to be reviewed by users (e.g. not behind a paywall)

Voting is then opened for a fortnight and advertised by social media. The ballots are then deduplicated (since some people accidentally vote more than once), results are tallied and released often with aggregate statistics shortly afterwards.

Criticisms and Concerns

Feedback has raised various criticisms and concerns over the years. Sometimes these misunderstand the nature of the activity as awareness-raising, but others are worth exploring. The most important criticism relates to language and culture – i.e. that the majority of resources found in DH Awards are in English or created in a Western context. Although DH Awards tries to counteract this with an ever-increasing international nomination committee, this criticism is certainly accurate. Trying to combat this in 2013 it introduced a 'Best DH contribution not in the English language' to encourage non-English submissions, but even more complaints rightly pointed out this ghettoized contributions, so it was removed the following year. DH Awards nominations have never been specific as to geography, language, conference, organization or field of humanities, however, it proactively seeks to encourage nominations from less represented areas and languages.

Other important issues raised include:

- The lack of awareness of their nature as community awards, with undue significance being ascribed to them in CVs, tenure portfolios, job applications, and funding bids, even though winning provides no financial value or prestige.
- Well-funded projects with a broad appeal are more likely to win than excellent smaller projects with a more specific remit.
- The annual nature of the awards ruling out older historic projects.

- The use of the optional demographic statistics being gathered.

This bulk of the paper will propose possible solutions to these concerns.

Bibliography

ADHO, 'Protocol for the Standing Committee on Awards of the Alliance of Digital Humanities Organizations', http://adho.org/administration/awards/protocol-standing-committee-awards-alliance-digital-humanities-organizations, Last Updated: 2008-06-24, Accessed: 2021-12-10

ADHO, 'Roberto Busa Prize', http://adho.org/awards/roberto-busa-prize, Accessed: 2021-11-10

Cummings, J. (2021) "DH Awards Frequently Asked Questions - DH Awards 2021", DH Awards, http://dhawards.org/dhawards2021/faqs/ Accessed: 2022-03-10

English, J. F. (2009) The Economy of Prestige: prizes, awards, and the circulation of cultural value, (Harvard University Press: 2009

Hockey, S. (2005), 'Living with Google: Perspectives on Humanities Computing and Digital Libraries: Busa Award Lecture', June 2004, Literary and Linguistic Computing, Volume 20, Issue 1, March 2005, 7–24, https://doi.org/10.1093/llc/fqh040 Accessed: 2021-12-10

Notes

1. See http://adho.org/awards/roberto-busa-prize for more information about the ADHO Roberto Busa Prize.
2. The protocols in 2004 weren't openly or at least widely available. At the time of writing, the protocol seems to have remained the same since 2008 and there are certainly some aspects that need updating. See http://adho.org/administration/awards/protocol-standing-committee-awards-alliance-digital-humanities-organizations for the protocol of the ADHO SCA. If the representatives of the committee had been approached in 2004, it is likely they would have been happy to explain the process.
3. See James F. English, The Economy of Prestige: prizes, awards, and the circulation of cultural value, (Harvard University Press: 2009) as well as many other discussions for more about the introduction of awards of various sorts as catalysts and awareness-raising into areas of cultural production. There is a strong influence on DH Awards' creation from cognate award systems such as those in film and web-based awards.

Emotion courses in German historical comedies and tragedies

Dennerlein, Katrin
katrin.dennerlein@uni-wuerzburg.de
Julius Maximilians University Würzburg, Germany, Germany

Schmidt, Thomas
thomas.schmidt@ur.de
Media Informatics Group, University of Regensburg, Germany

Wolff, Christian
christian.wolff@ur.de
Media Informatics Group, University of Regensburg, Germany

Introduction

From 1650 to early nineteenth century the drama in the German speaking area develops rapidly and turns into the most popular genre of this period (Brenner, 1999; Meid, 2009: 327-501). It becomes a 'school of affects' (Rotermund, 1972: 25). Until now, literary scholars have mostly investigated individual emotions, examining selected plays in detail (Schings, 1980; Meier, 1993; Schulz, 1988; Zeller 2005; Anz, 2011; Schonlau, 2017). As a result, little is known about which emotions play a role in character communication in specific genres for this period. In computational literary studies, emotional aspects in dramatic texts have been studied only sporadically in comparison with prose fiction (Kim and Klinger, 2019; Jacobs, 2019). Regarding plays, the main focus has been the analysis of valence or polarity, and mostly on individual authors or works (Mohammad, 2011; Nalisnick and Baird, 2013; Schmidt and Burghardt, 2018; Schmidt et al., 2019b; Schmidt et al., 2019c; Schmidt and Wolff, 2021). In this paper, we will present first results on the prediction of emotions in 226 comedies and tragedies from the 17th to the early 19th century using state-of-the-art language models (for more information see Schmidt et al., 2021a; 2021b; 2021c; Dennerlein et al., 2022b; Schmidt et al; 2022). This research is part of the project "Emotions in Drama". [1]

Emotion Set, Annotation

We define 'emotions' as internally represented and subjectively experienced categories that can be registered by the individual in an ego-related and introspective-mental as well as physical way. They may express themselves in perceptible variations of expression (Schwarz-Friesel, 2007: 55). We annotate intended emotions experienced by and attributed to characters. Following an extensive study of the affect theories of the time (Zeller 2005; Grimm 2010), we have worked out definitions that closely follow the historical concepts and have developed an annotation scheme with many examples and some further distinctions (Dennerlein et al. 2022a). We decided to annotate the following emotions:

- Emotions of affection / Zuneigung
  - desire / Lust (+)
  - love / Liebe (+)
  - friendship / Freundschaft (+)
  - admiration, reverence / Verehrung, Bewunderung (+)
- Emotions of pleasure / Freude und Glück
  - joy / Freude (+)
  - Schadenfreude (+)
- Emotions of anxiety / Angst und Sorge
  - fear / Angst (-)
  - despair / Verzweiflung (-)
- Emotions of rejection / Ablehnung
  - anger / Ärger (-)
  - abhorrence / Abscheu, Wut, Hass (-)
- Emotions of suffering and empathy / Leid
  - suffering / Leid (-)
  - compassion / Mitleid (-)
- no main class
  - being moved / emotionale Bewegtheit (undetermined)

The main criterion for the choice of emotion categories was that the selection should make it possible to represent changes in literary history and differences in genre. So far, these emotions have been annotated in 11 dramas (5 comedies and 6 tragedies from 1745-1807) by two independent annotators each resulting more than 13,000 annotations (Schmidt et al., 2021a). Annotators could annotate text spans of variable size ranging from one word to several sentences because the expression of emotions can refer to text segments of different lengths. The inter-annotator agreements range from 0.3 to 0.4 for κ-values at the emotion level, depending on the drama, which corresponds to a weak to moderate agreement (Landis and Koch, 1977). These comparatively low agreement values are common for the annotation of historical and literary texts (Alm and Sproat, 2005; Sprugnoli et al., 2016; Schmidt et al., 2018; Schmidt et al., 2019a; Schmidt et al., 2019; Schmidt et al., 2020). We intend to further enhance the scores through continuous improvement of the annotation guidance and training of the annotators.

Computational Emotion Classification

We evaluated multiple computational single-label classification approaches on the emotion annotations for the emotion classification of the 13 emotions and (polarity) classes (Schmidt et al., 2021a; 2021c). The highest accuracies were achieved with the transformer-based model *gbert* by *deepset* (Chan et al., 2020) finetuned to the emotion classification task with all annotations filtered by the disagreements of the two annotators (resulting in 7,000–10,000 annotations depending on the hierarchical system). This model achieves accuracies ranging from 90% (polarity) to 67% (sub-emotions) and outperforms the more commonly used method of lexicon-based sentiment analysis in DH (Kim and Klinger, 2019; Fehle et al., 2021). More information about the results and the model can be found in Schmidt et al. (2021c). The computational emotion classification used in the next parts was applied on the sentences of the plays for 123 comedies and 103 tragedies from 1650-1829.

Emotions in comedies and tragedies: annotation vs. classification

We analyze the frequency of emotion annotations and the computational classifications throughout the plot of the drama. For that goal each drama is divided into five equal parts (quintiles) because it allows for normalized comparisons. We calculated the average number of annotations (for 11 plays) and computational emotion classifications (for 226 plays) per quintile for each genre.

Fig. 1 shows the distribution of the emotion 'suffering' in the plot of the annotated dramas. The emotion was annotated on average exactly twice as often in tragedies as in comedies (on average 27-32 passages with suffering in the comedy, 45-60 in the tragedy).

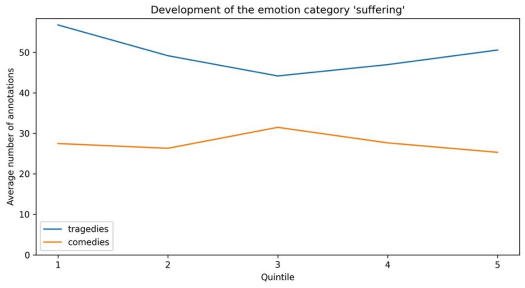


Fig. 1:

Development of 'suffering' as measured in annotations for tragedies and comedies.

Moreover, one can see in fig. 1 that suffering is clustered at the beginning and end of tragedies. In the middle of the tragedies, however, there is obviously hope for an improvement of the situation and the characters feel less suffering. In comedies, on the other hand, after a brief decrease in suffering, we recognize a suffering climax, which can be interpreted as the turning point towards a good ending. In fig. 2, we visualize the average amount of sentences classified as suffering by the computational emotion classification throughout the 5 quintiles of the plays. Fig. 3 illustrates the opposite emotion joy for the 1,619 annotations in the annotated plays.

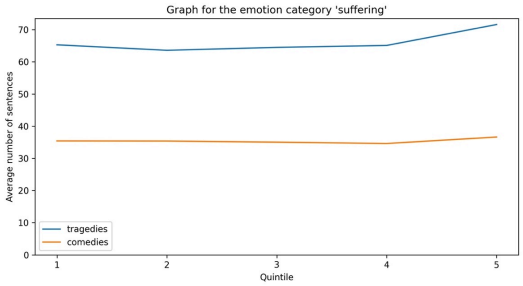


Fig. 2:
Development of 'suffering' as measured by the emotion classification in tragedies and comedies.

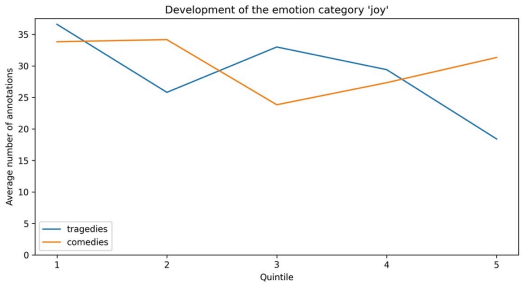


Fig. 3:
Development of 'joy' as measured by the annotations in tragedies and comedies.

In comedies, joy is least annotated in the middle of the plot, when confusion and problems accumulate; towards the end, the values rise again almost to the level of the beginning (fig. 3). In tragedies, on the other hand, the most joyful statements by characters are found shortly before the middle of the plot (fig. 3). This finding of a sudden drop in joy correlates with the dramaturgical concept of *peripetia*, the change of happiness. According to the ideal-typical Aristotelian definition, the change of action inevitably leads to a bad ending. The results of our annotation analysis show a matching steady decline of joy in tragedy.

Fig. 4, however, shows that the emotion classification model produces different results.

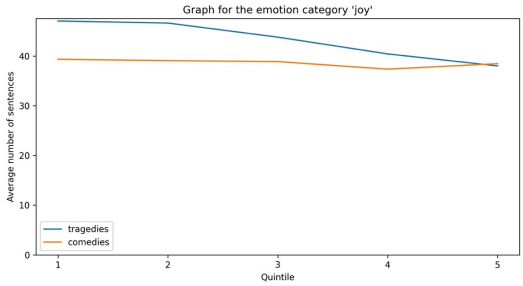


Fig. 4:
Development of 'joy' as measured by the emotion classification in tragedies and comedies.

Particularly interesting in fig. 4 is the fact that the absolute number of joy sentences is higher in the tragedies than in the comedies. However, it is clear that joy then decreases much more in tragedies than in comedies, which increases the tragic effect of the tragedies. Compared to the annotated comedies, the curve for joy in the comedies shows little change. In our presentation, we will discuss whether we are dealing with the larger deviations between figs. 3 and 4 as an indication of the still insufficient quality of the prediction, or whether the results are rather related to the specific tragedy and comedy subgenres that predominate in our corpus and that have less ideal-typical developments than the annotated dramas.

Bibliography

Alm, C. O. and Sproat, R. (2005). Emotional Sequencing and Development in Fairy Tales. In Tao, J., Tan, T. and Picard, R. W. (eds), *Affective Computing and Intelligent Interaction*. (Lecture Notes in Computer Science). Berlin, Heidelberg: Springer, pp. 668–74 doi: 10.1007/11573548_86.

Anz, T. (2011). Todesszenarien: Literarische Techniken zur Evokation von Angst, Trauer und anderen Gefühlen. In Ebert, L. (ed), *Emotionale Grenzgänge. Konzeptualisierungen von Liebe, Trauer und Angst in Sprache und Literatur*. Würzburg: Königshausen & Neumann, pp. 113–29.

Brenner, P. J. and Grimminger, Rolf (1999). Das Drama. *Die Literatur Des 17. Jahrhunderts*, vol. 2. (Hansers Sozialgeschichte Der Deutschen Literatur Vom 16.

Jahrhundert Bis Zur Gegenwart.). München/Wien, pp. 539–74.

Chan, B., Schweter, S. and Möller, T. (2020). German's Next Language Model. *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 6788–96 doi: 10.18653/v1/2020.coling-main.598. https://aclanthology.org/2020.coling-main.598 (accessed 15 February 2022).

Dennerlein, K., Schmidt, T. and Wolff, C. (2022a). Figurenemotionen in deutschsprachigen Dramen annotieren. Zenodo doi: 10.5281/zenodo.6228151. https://zenodo.org/record/6228152 (accessed 21 April 2022).

Dennerlein, K., Schmidt, T. and Wolff, C. (2022b). Emotionen im kulturellen Gedächtnis bewahren. *DHd 2022 Kulturen des Digitalen Gedächtnisses. 8. Tagung des Verbands 'Digital Humanities im Deutschsprachigen Raum' (DHd 2022)*. Potsdam, Germany: Zenodo, pp. 93–98 doi: 10.5281/zenodo.6327957. https://zenodo.org/record/6327957 (accessed 21 April 2022).

Fehle, J., Schmidt, T. and Wolff, C. (2021). Lexicon-based Sentiment Analysis in German: Systematic Evaluation of Resources and Preprocessing Techniques. *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*. Düsseldorf, Germany: KONVENS 2021 Organizers, pp. 86–103 https://aclanthology.org/2021.konvens-1.8 (accessed 21 April 2022).

Grimm, H. (1980). Affekt. In Barck, K., Fontius, M., Schlenstedt, D., Burkhart, S. and Wolfzettel, F. (eds), *Ästhetische Grundbegriffe*, vol. 1. pp. 16–49.

Kim, E. and Klinger, R. (2019). A Survey on Sentiment and Emotion Analysis for Computational Literary Studies. *Zeitschrift Für Digitale Geisteswissenschaften*. Herzog August Bibliothek doi: 10.17175/2019_008_V2. https://zfdg.de/2019_008 (accessed 14 February 2022).

Landis, J. R. and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *International Biometric Society*, **33**(1). [Wiley, International Biometric Society]: 159–74 doi: 10.2307/2529310.

Meid, Volker (2009). *Die Deutsche Literatur Im Zeitalter des Barock. Vom Späthumanismus zur Frühaufklärung 1570–1740*. (Ed.) De Boor, Helmut & Newald, Richard (Geschichte der Deutschen Literatur von der Aufklärung bis zur Gegenwart). München: Beck.

Mohammad, S. (2011). From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Portland, OR, USA: Association for Computational Linguistics, pp. 105–14 https://www.aclweb.org/anthology/W11-1514 (accessed 21 March 2021).

Nalisnick, E. T. and Baird, H. S. (2013). Character-to-Character Sentiment Analysis in Shakespeare's Plays. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 479–83 https://www.aclweb.org/anthology/P13-2085 (accessed 1 May 2020).

Rotermund, E. (1972). *Affekt und Artistik: Studien zur Leidenschaftsdarstellung und zum Argumentationsverfahren bei Hofmann von Hofmannswaldau*. (7). München: W. Fink.

Schings, H.-J. (1980). *Der Mitleidigste Mensch ist der Beste Mensch: Poetik des Mitleids von Lessing bis Büchner*. (Edition Beck). München: Beck.

Schmidt, T. and Burghardt, M. (2018). An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing. *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Santa Fe, New Mexico: Association for Computational Linguistics, pp. 139–49 https://www.aclweb.org/anthology/W18-4516 (accessed 6 April 2020).

Schmidt, T., Burghardt, M. and Dennerlein, K. (2018). Sentiment Annotation of Historic German Plays: An Empirical Study on Annotation Behavior. In Kübler, S. and Zinsmeister, H. (eds), *Proceedings of the Workshop on Annotation in Digital Humanities (AnnDH 2018)*. Sofia, Bulgaria, pp. 47–52 http://ceur-ws.org/Vol-2155/schmidt.pdf (accessed 20 April 2022).

Schmidt, T., Burghardt, M., Dennerlein, K. and Wolff, C. (2019a). Katharsis – A Tool for Computational Drametrics. *Book of Abstracts, Digital Humanities Conference 2019 (DH 2019)*. Utrecht, Netherlands https://dev.clariah.nl/files/dh2019/boa/0584.html (accessed 23 May 2021).

Schmidt, T., Burghardt, M., Dennerlein, K. and Wolff, C. (2019b). Sentiment Annotation for Lessing's Plays: Towards a Language Resource for Sentiment Analysis on German Literary Texts. In Declerck, T. and McCrae, J. P. (eds), *Proceedings of the Poster Session of the 2nd Conference on Language, Data and Knowledge (LDK-PS 2019)*. Leipzig, Germany: RWTH Aachen, pp. 45–50 http://ceur-ws.org/Vol-2402/paper9.pdf (accessed 21 April 2022).

Schmidt, T., Burghardt, M. and Wolff, C. (2019c). Toward Multimodal Sentiment Analysis of Historic Plays: A Case Study with Text and Audio for Lessing's Emilia Galotti. In Navarretta, C., Agirrezabal, M. and Maegaard, B. (eds), *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference (DHN 2019)*.

Copenhagen, Denmark, pp. 405–14 http://ceur-ws.org/Vol-2364/37_paper.pdf (accessed 21 April 2022).

Schmidt, T., Dennerlein, K. and Wolff, C. (2021a). Towards a Corpus of Historical German Plays with Emotion Annotations. *3rd Conference on Language, Data and Knowledge (LDK 2021)*, vol. 93. (Open Access Series in Informatics (OASIcs)). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, p. 9:1-9:11 doi: 10.4230/OASIcs.LDK.2021.9. https://drops.dagstuhl.de/opus/volltexte/2021/14545 (accessed 21 April 2022).

Schmidt, T., Dennerlein, K. and Wolff, C. (2021b). Using Deep Learning for Emotion Analysis of 18th and 19th Century German Plays. In Burghardt, M., Dieckmann, L., Steyer, T., Trilcke, P., Walkowski, N.-O., Weis, J. and Wuttke, U. (eds), *Fabrikation von Erkenntnis: Experimente in Den Digital Humanities.* Esch-sur-Alzette, Luxembourg: Melusina Press doi: https://www.melusinapress.lu/read/10-26298-melusina-8f8w-y749-udlf/section/8d0fefff-384c-4798-b5d7-032809de2430 (accessed 20 April 2022).

Schmidt, T., Dennerlein, K. and Wolff, C. (2021c). Emotion Classification in German Plays with Transformer-based Language Models Pretrained on Historical and Contemporary Language. *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Punta Cana, Dominican Republic (online): Association for Computational Linguistics, pp. 67–79 doi: 10.18653/v1/2021.latechclfl-1.8. https://aclanthology.org/2021.latechclfl-1.8 (accessed 21 April 2022).

Schmidt, T., Dennerlein, K. and Wolff, C. (2022). Evaluation computergestützter Verfahren der Emotionsklassifikation für deutschsprachige Dramen um 1800. *DHd 2022 Kulturen Des Digitalen Gedächtnisses. 8. Tagung Des Verbands 'Digital Humanities Im Deutschsprachigen Raum' (DHd 2022)*. Potsdam, Germany: Zenodo doi: 10.5281/zenodo.6328169. https://zenodo.org/record/6328169 (accessed 21 April 2022).

Schmidt, T., Engl, I., Halbhuber, D. and Wolff, C. (2020). Comparing Live Sentiment Annotation of Movies via Arduino and a Slider with Textual Annotation of Subtitles. In Reinsone, S., Skadiņa, I., Daugavietis, J. and Baklāne, A. (eds), *Post-Proceedings of the 5th Conference Digital Humanities in the Nordic Countries (DHN 2020)*, vol. 2865. Riga, Latvia: CEUR Workshop Proceedings, pp. 212–23 http://ceur-ws.org/Vol-2865/poster1.pdf (accessed 21 April 2022).

Schmidt, T., Winterl, B., Maul, M., Schark, A., Vlad, A. and Wolff, C. (2019d). Inter-Rater Agreement and Usability: A Comparative Evaluation of Annotation Tools for Sentiment Annotation. In Draude, C., Lange, M. and Sick, B. (eds), *INFORMATIK 2019: 50 Jahre Gesellschaft Für Informatik – Informatik Für Gesellschaft (Workshop-Beiträge)*. Bonn: Gesellschaft für Informatik e.V., pp. 121–33 doi: 10.18420/inf2019_ws12.

Schmidt, T. and Wolff, C. (2021). Exploring Multimodal Sentiment Analysis in Plays: A Case Study for a Theater Recording of Emilia Galotti. *Proceedings of the Conference on Computational Humanities Research 2021 (CHR 2021)*. Amsterdam, The Netherlands, pp. 392–404.

Schonlau, Anja (2017). *Emotionen im Dramentext: Eine Methodische Grundlegung mit Exemplarischer Analyse zu Neid und Intrige 1750-1800*. Berlin, Boston: De Gruyter.

Schulz, G.-M. (1988). *Tugend, Gewalt und Tod: das Trauerspiel der Aufklärung und die Dramaturgie des Pathetischen und des Erhabenen*. Tübingen: Niemeyer.

Schwarz-Friesel, M. (2007). *Sprache und Emotion*. Tübingen: Francke.

Sprugnoli, R., Tonelli, S., Marchetti, A. and Moretti, G. (2015). Towards Sentiment Analysis for Historical Texts. *Digital Scholarship in the Humanities*, **31**. Oxford: Oxford University Press: 762–72 doi: 10.1093/llc/fqv027.

Zeller, R. (2005). Tragödientheorie, Tragödienpraxis und Leidenschaften. In Steiger, J. A. (ed), *Passion, Affekt und Leidenschaft in der Frühen Neuzeit*, vol. II. Wiesbaden: Harrassowitz, pp. 691–704.

Notes

1. The project "Emotions in Drama" (EmoDrama) is funded by the DFG (German Research Association) in the priority program Computational Literary Studies (SPP 2207/1) with two grants (project number 424207618; grants DE 2188/3-1 und WO 835/4-1). For more information see https://dfg-spp-cls.github.io/projects_en/2020/01/24/TP-Emotions_in_Drama/

Database Design and Identity: A Compromised Infrastructure

Earhart, Amy
aearhart@tamu.edu
Texas A&M University, United States of America

This paper will address issues of representing identity in databases, with a particular focus on Asian identities. Identity, such as that the conference theme defines, is often far more diverse and fluid than what a binary technology tool might represent. Utilizing the Database of African American and Predominantly White American Literature Anthologies or DALA, a database of 100 years of American

and African American literature anthologies, constructed to investigate questions of identity and representation, this paper calls for theory-based, historically aware, transparent approaches to encoding identity in dh tools like the database.

Dh projects are increasingly interested in representing identity across time, whether through analysis of census data, historical records, or literary history. Much of this work is completed using databases, with human bodies turned into binary data points contained within rigid columns, a method that is resistance to contemporary theorization of identity. This project is informed by the growing wealth of scholarship produced by algorithmic bias scholars who have been attentive to the nuances of race and gender within technological infrastructures (Noble 2018, Benjamin 2019, McIlwain 2020, Brock 2020, Steele 2021, Womack 2022?), working across issues from social justice, algorithmic bias, surveillance, and social media. Information studies also offers analysis that we might use to address bias (Drabinski 2013, Block 2020). While DAAPWALA provides interesting results, it is, in many ways, designed as a project to test how a database might use categories of identity in transparent ways, modeling both problems and possibilities within a system that is insistent upon the binary. As Tara McPherson has argued, the best digital database projects: "…wield technology against its positivist self, foregrounding the work of the interface and refusing an easy transparency and corporate tenents of 'good' design via the template" (2015, 495).? It is this uneasiness that I hope to explore in the paper, designed to begin a conversation within the digital humanities community about best practices for encoding identity.

To interpret canonicity and inclusion/exclusion of authors in literary canons using a database, authors and editors must be categorized in a manner that is consistently understood throughout the body of the data, making static any categories into which individuals are encoded. For example, the database follows "good" database practices in designing the database, encoding designated vocabulary in drop down menus to regularize the selection of everything from author names to identities, adding viaf numbers, an international authority file, to further control identity and ensure interoperability. Such methods do regularize the data, something that needed to happen for comparative purposes and to insure consistency, particularly as there were numerous individuals inputting data over time, but the encoding methods are opposed to contemporary theoretical models of identity, a conundrum that will be discussed further.

The paper will discuss the choices made in identifying authors and editors as "Asian." As the call for papers notes, dh has a long history in Asia. The conference theme is designed to center Asian dh scholars and their work, resisting the displacement of Asian dh work, as is often the case at ADHO conferences held in Europe and the Americas. A similar representation of identity, decentering western whiteness, is central to encoding identity within DAAPWALA. For the purposes of DAAPWALA I use the term Asian to include East Asian, South Asian, Southeast Asian, Central Asian and Pacific Islander. I will discuss the historical and cultural components that are used for categorizations, highlighting best practices used in TEI/XML encoding and library metadata standards. I argue that while we must stress interoperability, we must also account for the development of data as a representation of the local environment. In the case of the United States, and the literary anthologies I am studying, identity terms are not developed in a vacuum. Historic and contemporary racism and government policy impacted identification of groups that slip in and out of identities. I reject the use of census categories as the long history of the U.S. census is problematic and doesn't meet the goals of DAAPWALA. I will also discuss the importance of working across national and cultural divides while also attending to the particularities of the local environment. Asian as define defined in Asian American literary anthologies does not have the same cultural markers as Asian as defined by the conference call. Further, I will discuss the importance of data base design that considers the differing ways that cultures regularize identity markers, such as names. While databases developed in the United States, for example, use a field for first name and a field for last name this is not consistent across all cultures and broader naming conventions are needed. Emphasizing the need to balance interoperability against cultural specificity, this paper asks the dh community to begin conversations about better data design.

Bibliography

Benjamin, R. (2019). Race After Technology: Abolitionist Tools for New Jim Code. Cambridge: Polity Press.

Block, S. (2020). Erasure, Misrepresentation and Confusion: Investigating JSTOR Topics on Women's and Race Histories. DHQ: Digital Humanities Quarterly 14, no. 1.

Brock, A. (2020.) Distributed Blackness: African American Cybercultures. New York: New York University Press.

Drabinski, E. (2013). "Queering the Catalog: Queer Theory and the Politics of Correction." The Library Quarterly 83, no. 2: pp. 94–111.

Gallon, K. (2016.) "Making a Case for the Black Digital Humanities." In Debates in the Digital Humanities 2016, Matthew K. Gold and Lauren F. Klein (eds.). Minneapolis: U Minnesota P.

McIlwain, C. (2020). Black Software The Internet & Racial Justice, from the AfroNet to Black Lives Matter. New York: Oxford UP.

McPherson, T. (2015). "Post-Archive: The Humanities, the Archive, and the Database." In Between Humanities and the Digital, Patrik Svensson and David Theo Goldberg (eds.) pp. 483–502. Cambridge, Mass: MIT Press.

Noble, SU. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism. New York: NYU Press.

Steele, CK. (2021). Digital Black Feminism. New York: NYU Press.

Womack, A. (2022) The Matter of Black Living: The Aesthetic Experiment of Racial Data, 1880–1930. Chicago: U of Chicago P.

One word to rule them all: understanding word embeddings for authorship attribution

Eder, Maciej
maciej.eder@ijp.pan.pl
Institute of Polish Language (Polish Academy of Sciences)

Šeļa, Artjoms
atrjoms.sela@ijp.pan.pl
Institute of Polish Language (Polish Academy of Sciences)

Introduction

With an advent of deep learning in natural language processing, the ways in which a text could be represented became much more complex and much less transparent. From simple estimations of word frequency distributions, methods shifted to context-aware embeddings and neural network generalizations. These opaque representations made their way to the authorship attribution with obvious improvements across different tasks (Benzebouchi et al. 2018; Gómez-Adorno et al. 2018; Kiros et al. 2014; Posadas-Durán et al. 2017). Yet, this improvement did not bring us closer to understanding of authorial style. Rather the opposite happened, and we have obscured the reasons for feature effectiveness in attribution tasks.

In this study we want to ask how much of the contextual information matters for authorship attribution in a conservative setting (authors represented by a few large fictional texts). We propose a simple experimental setup with a basic word embedding model that represents words by their contexts (or co-occurrence with other words in immediate proximity); in such a setup, each word is represented by a vector of coordinates in a semantic space, instead of using traditional word frequencies. In order to get to a text-wide vector representation, we use several tactics, derived from sentence/paragraph embedding approach (Le & Mikolov 2014) and is similar to "timestamping" tokens (Dubossarsky et al. 2019). Our approach mostly boils down to adding quasi-tokens that are tied to each specific text in the corpus: they act as sponges, soaking word occurrences from their immediate context. We use this sponges as "text embeddings" that share the same vector space with actual words from the corpus. Having control on text and context representation, we manipulate the underlying words (by randomly shuffling them or changing the principles of quasi-tokens distributions).

	D_1	D_2	D_3	D_4	D_5	…
morning	0.402	0.716	–0.930	–0.264	–0.046	…
the_Barclay_1	0.469	0.351	0.054	–0.979	0.171	…
table	–0.810	0.255	–0.675	0.227	1.059	…
breakfast	–1.010	0.485	-0.542	0.462	0.500	…
the_Bennet_2	–0.072	0.295	-0.212	–0.640	1.020	…
…	…	…	…	…	…	…

What we learn, is that accurate contextual representation does not matter for attribution task: text embeddings of randomized novels work similarly to embeddings that preserve original word order and learn context-aware semantic representations. In other words, we can have accurate authorship classification even if individual word vectors and their contextual neighborhoods contain but noise. This suggests that the authorship signal continues to be largely driven by underlying document-specific distributions of word frequencies.

What we learn, is that accurate contextual representation does not matter for attribution task: text embeddings of randomized novels work similarly to embeddings that preserve original word order and learn context-aware semantic representations. In other words, we can have accurate authorship classification even if individual word vectors and their contextual neighborhoods contain but noise. This suggests that the authorship signal continues to be largely driven by underlying document-specific distributions of word frequencies.

Data and methods

We use "100 English novels" as our test corpus, which was employed in other benchmarking experiments (Rybicki & Eder 2011; Eder 2013). It has 33 authors represented by

3 novels each. On the one hand, the authorship recognition is a rather trivial task in this case, since chronological distribution of authors is wide (from Bronte sisters to Virginia Woolf). On the other hand, the attribution scenario is not that trivial, since the classifier has to search through 33 candidate classes for the correct answer. Despite possible limitations of our benchmark dataset, however, we consider it to be suitable for experimental setups that do not try to "stress-test" methods or claim improvements over state-of-the-art.

Over the course of experiments we use GloVe algorithm for learning word association using matrix decomposition, without any shallow or deep neural networks. We prepare text representation within a word-based model in following ways:

- MFW-sponge. Given a word *X* from the list of 200 most frequent words, we transform all *X* tokens by adding to them their corresponding text IDs. E.g.: "the_Doyle_1". We assume that tying identity tokens to frequent words gives a natural access to wide contexts in which these function words occur.
- Dummy-sponge. Instead of learning existing tokens, we add non-existent identity tokens randomly to each text. E.g.: "Mr. DUMMY_Doyle_1 Sherlock Holmes, who DUMMY_Doyle_1 was".

The frequency of dummy sponges is also determined by several approaches

- Dummy-token frequency (DTF) is tied to frequency of "the" in a given text;
- DTFis taken randomly from normal distribution with μ equal to the mean relative frequency of the word "the" across the corpus, and σ equal to the standard deviation of "the";
- DTF is constant, based on max relative frequency of "the" in the corpus;
- DTF is an arbitrary constant, larger than the most frequent word (we have picked the value of 0.08);
- DTF is inferred through a *word-rank ~ frequency* model (what would be a frequency of a word more frequent than most frequent word in a natural language?).

We apply this text-embedding logic to 1) original text; 2) randomly shuffled text; 3) text where all words under a certain frequency threshold were replaced. We train small-scale GloVe models for each variation in the textual setup (100 novels are the only source of learning the contexts).

For authorship attribution, we use SVM classification with linear kernel, using "identity vectors" or sponges as text representation. Each quasi-token has the same dimensionality as the original GloVe model (300 in our case). Importantly, we test one sponge at a time, using its 300 dimensions as features for SVM. We cross-validate each model randomly placing 2 novels for each author to represent test set and leaving 1 (33 novels altogether) for the training set.

Results

First and foremost, the identity vector approach works at a competitive level. The inter-textual relationships could be roughly represented via UMAP (Fig. 1) that projects all text-vectors alongside with the subset of the most similar words (cosine similarity). We achieve 0.93 median accuracy, which is slightly better than the methodological baseline that uses simple MFW approach.

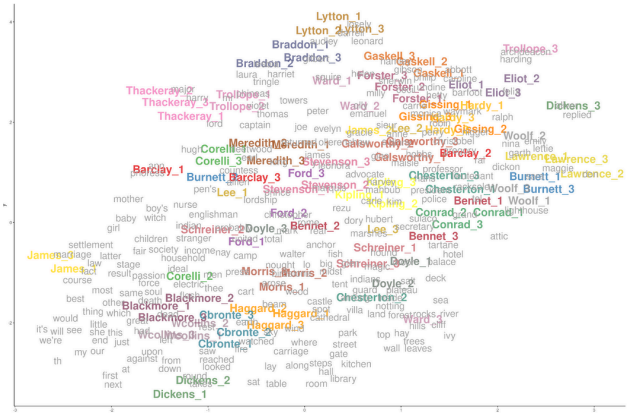


Fig. 1:
UMAP projection of the sponge-vectors "the" with their 20 immediate contextual neighbors (cosine similarity, repeated neighbors were represented only once).

Secondly, we notice that performance is strongly related to the frequencies of quasi-tokens (Fig. 2), suggesting that the more access to parts of target text we have, the better. However, the position of the tokens themselves did not matter: best dummy-sponge performance was at the level of best MFW-sponges performance.

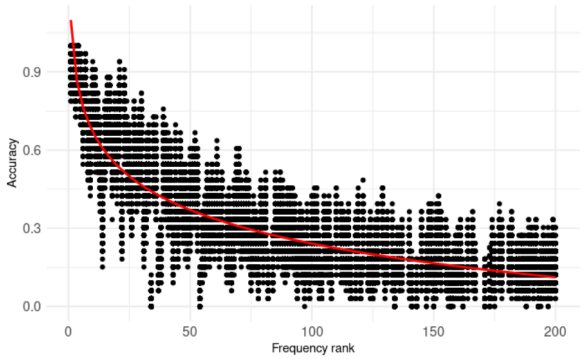


Fig. 2:
Linear model that predicts decrease in accuracy with increase in log-frequency rank. Increase in each log-unit in quasi-token frequency rank results in ~18% accuracy drop. Points are showing original accuracy scores from cross-validation

Thirdly, it turned out that the position of tokens didn't affect the classification results, and the same could be said about the original structure of the text, suggesting that the word order (and thus the context) is neglectable for attribution. If all the words across all the novels were randomly shuffled, the performance of dummy-sponges stayed at the same level. Representation of word semantics became meaningless, but the attribution task was still brute-forced by the information retained by the sponges.

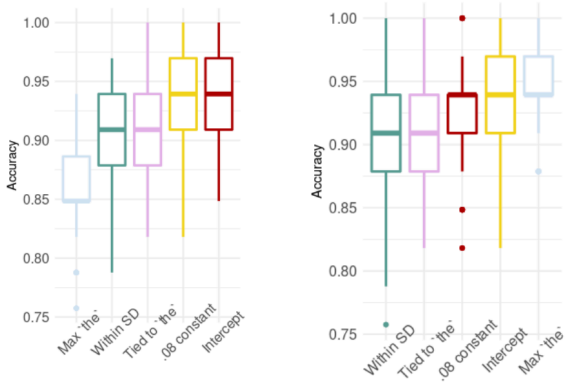


Fig. 3:
Results for dummy-sponges in original texts (left). Results for dummy sponges that "soak context" in randomly shuffled pools of words (right).

Furthermore, in a scenario when only 500 most frequent words were left in the original texts, while the remaining words were replaced by flat fillers (e.g. "DUMMY DUMMY DUMMY was DUMMY in the DUMMY DUMMY"), our sponge approach was still able to produce accurate attributions (albeit worse, with median accuracy at 0.81).

Discussion

Our results show that it is possible to completely remove context-dependent semantics from texts, yet embedded text-wide vectors will still perform well for authorship attribution tasks. This strongly suggests that contextual representation in authorship attribution remains driven by the sheer frequency of the most frequent units of language, and the chance for identity tokens to be exposed to text-specific surroundings. This not only highlights that complex and simple text representations draw from the same source of word frequency distributions (cf. Dębowski 2021), but also suggests the path to formal independent modeling of style and meaning, allowing for complex style transfers and adversarial stylometry (Brennan et al. 2012), where the task is to mask the style of a writing, but preserve the message for maintaining anonymity.

Acknowledgements

This research is part of the project 2017/26/E/HS2/01019, supported by Poland's National Science Centre.

Bibliography

Benzebouchi, N.E., Azizi, N., Aldwairi, M., Farah, N., 2018. Multi-classifier system for authorship verification task using word embeddings. In: *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*. Presented at the 2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP), pp. 1–6. https://doi.org/10.1109/ICNLSP.2018.8374391

Brennan, M., Afroz, S., Greenstadt, R., 2012. Adversarial Stylometry: Circumventing Authorship Recognition to Preserve Privacy and Anonymity. *ACM Trans. Inf. Syst. Secur.* 15. https://doi.org/10.1145/2382448.2382450

Dębowski, Ł. 2021. A Refutation of Finite-State Language Models through Zipf's Law for Factual Knowledge. *Entropy* 23(9): 1148. https://doi.org/10.3390/e23091148

Dubossarsky, H., Hengchen, S., Tahmasebi, N., Schlechtweg, D., 2019. Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Presented at the ACL 2019,

Association for Computational Linguistics, Florence, Italy, pp. 457–470. https://doi.org/10.18653/v1/P19-1044

Eder, M., 2013. Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities* 30, 167–182. https://doi.org/10.1093/llc/fqt066

Gómez-Adorno, H., Posadas-Durán, J.-P., Sidorov, G., Pinto, D., 2018. Document embeddings learned on various types of n-grams for cross-topic authorship attribution. *Computing* 100, 741–756. https://doi.org/10.1007/s00607-018-0587-8

Kiros, R., Zemel, R.S., Salakhutdinov, R., 2014. A Multiplicative Model for Learning Distributed Text-Based Attribute Representations, in: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS'14*. MIT Press, Cambridge, MA, USA, pp. 2348–2356.

Rybicki, J., Eder, M., 2011. Deeper Delta across genres and languages: do we really need the most frequent words? *Literary and Linguistic Computing* 26, 315–321. https://doi.org/10.1093/llc/fqr031

Le, Q., Mikolov, T., 2014. Distributed Representations of Sentences and Documents, in: *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32, ICML'14*. JMLR.org, p. II-1188-II–1196.

Posadas-Durán, J.-P., Gómez-Adorno, H., Sidorov, G., Batyrshin, I., Pinto, D., Chanona-Hernández, L., 2017. Application of the distributed document representation in the authorship attribution task for small corpora. *Soft Comput* 21, 627–639. https://doi.org/10.1007/s00500-016-2446-x

Measuring Keyness

Evert, Stephanie
stephanie.evert@fau.de
FAU Erlangen-Nürnberg, Germany

Keywords in corpus linguistics and DH

In corpus linguistics, the notion of **keywords** refers to words (and sometimes also multiword units, semantic categories or lexico-grammatical constructions) that "occur with unusual frequency in a given text" (Scott, 1997: 236) or a text collection, i.e. a corpus. Keywords are deemed to represent the characteristic vocabulary of the target text or corpus and thus have many applications in corpus linguistics, digital humanities and computational social science. They can capture the aboutness of a text (Scott, 1997), the terminology of a text genre or technical domain (Paquot and Bestgen, 2009), important aspects of literary style (Culpeper, 2009), linguistic and cultural differences (Oakes and Farrow, 2006), etc.; they give insight into historical perspectives (Fidler and Cvrček, 2015) and provide a basis for measuring the similarity of text collections (Rayson and Garside, 2000). Keywords are also an important starting point for corpus-based discourse analysis (Baker, 2006), where manually formed clusters of keywords represent central topics, actors, metaphors, and framings (e.g. McEnery et al., 2015). Since this process is guided from the outset by human understanding, it provides a more interpretable alternative to topic models in hermeneutic text analysis.

Keywords are usually operationalised in terms of a statistical frequency comparison between the **target corpus** and a **reference corpus**. Different research questions can be addressed depending on the particular constellation of target T and reference R, e.g. (i) T = a single text vs. R = a text collection (➡ aboutness), (ii) T and R = collections of articles on the same topic in left-leaning and right-leaning newspapers (➡ contrastive framings), or (iii) T = texts from a given domain or genre vs. R = a large general-language reference corpus (➡ terminology).

Although keyword analysis is a well-established approach and has been implemented in many standard corpus-linguistic software tools such as WordSmith [1], AntConc [2], SketchEngine [3], and CQPweb (Hardie, 2012), it is still unclear what the "right" way of measuring keyness is (see overview in Hardie, 2014). In this paper, I propose (i) a mathematically well-founded **best-practice technique** and (ii) introduce a **visual approach** for exploring the empirical properties of different keyness measures.

Keyness measures

Keyword analysis is operationalised as a comparison of relative frequencies: For each **candidate** word, its frequency f_1 in a target corpus T of n_1 tokens is compared to its frequency f_2 in a reference corpus R of n_2 tokens. The candidate set of m items typically includes words that only occur in the target corpus ($f_2 = 0$).

A candidate is considered a ("positive") keyword if its relative frequency $p_1 = f_1 / n_1$ in T is substantially higher than its relative frequency $p_2 = f_2 / n_2$ in R. A large number of **keyness measures** have been proposed to quantify the comparison and thus provide a basis for a ranking of the candidates and/or cut-off thresholds. Three main groups of measures can be distinguished:

1. Measures based on **hypothesis tests** put the focus on establishing a statistically significant difference between p_1 and p_2. The most widely-used measures are chi-squared X^2 and log-likelihood G^2 (Dunning, 1993). These measures are biased towards high-frequency keywords, often including function words and other non-specific words.
2. **Effect size** measures instead focus on how many times more frequent a candidate is in T than in R. The most intuitive measure is relative risk $r = p_1 / p_2$, also known as LogRatio = $\log_2 r$ (Hardie, 2014). Some other effect-size measures are equivalent (%DIFF, Gabrielatos and Marchi, 2012) or closely related (odds ratio, Pojanapunya and Watson Todd, 2018) to LogRatio. These measures are biased towards very low-frequency keywords and are often combined with an additional significance filter (typically based on G^2).
3. Various **heuristic** measures lack any statistical foundation. They are often particularly easy to compute such as SketchEngine's SimpleMaths (Kilgarriff, 2009), which also offers a user parameter to adjust its bias towards high-frequency or low-frequency keywords.

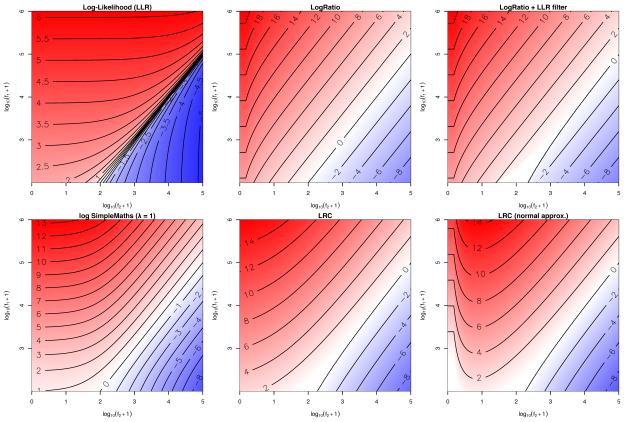
Mathematical discussion and visualisation

Hypothesis-test measures are subject to the criticism raised more generally against p-value testing in corpus linguistics and other fields (e.g. Gries, 2005). In particular, they are biased towards high-frequency keywords irrespective of effect size, selecting candidates that are not very salient for the target corpus. When they are applied more reasonably as a significance filter, the problem of multiple testing is often ignored: a single analysis may carry out frequency comparisons for hundreds of thousands of candidates, resulting in large numbers of false positives at customary significance levels such as $p < .001$ (Gries, 2005; Hardie, 2014).

By contrast, effect-size measures such as LogRatio are biased towards low-frequency keywords because they completely ignore the statistical significance of the observed difference in relative frequency. Moreover, many of these measures are undefined for $f_2 = 0$ and need special heuristics for this case; e.g. Hardie (2014) simply substitutes $f_2 = 0.5$ without mathematical justification.

Traditionally, keyness measures are computed from cumulative token frequency counts for T and R. However, two recent studies have independently concluded that keywords based on document counts are more robust (Evert et al., 2018; Egbert and Biber, 2019).

Keyness measures can also be understood from a more intuitive angle by visualising them as **topographic maps**, which show the scores assigned to all possible combinations of frequencies f_1 in T and f_2 in R on a logarithmic scale (similar to the visualisation of collocations in Evert, 2004: sec. 3.3). The examples in Fig. 1 reveal the respective frequency biases of G^2 and LogRatio – which is hardly mitigated by an additional significance filter – in the top row (dark red colours indicate frequency profiles of highly-ranked keywords).



Visualisation of keyness measures as topographic maps for $n_1 = n_2 = 100$ M words. The bottom right panel highlights problems of an earlier version of LRC currently used by CQPweb.

Best-practice recommendation

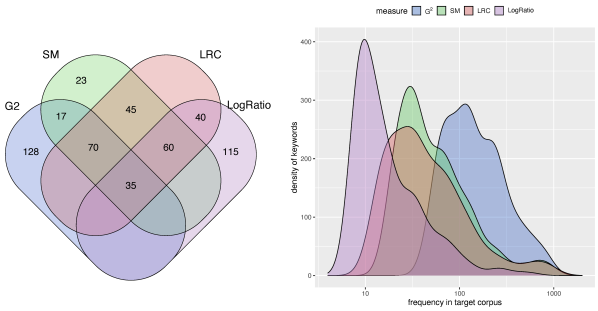
Conservative estimates based on statistical confidence intervals combine the advantages of hypothesis tests and effect-size measures into a single score. I therefore propose **LRC**, a conservative estimate of LogRatio, as a best-practice keyness measure. LRC uses an exact conditional Poisson test (Fay, 2010: 55) to obtain reliable confidence intervals corrected for multiple testing. The full procedure for computing LRC scores is as follows:

1. Collect the frequency data f_1, f_2 for each candidate and the sample sizes n_1, n_2 of T and R. Wherever suitable, document frequencies should be preferred.
2. Compute a two-sided Pearson-Clopper binomial confidence interval $[\pi_-, \pi_+]$ for f_1 successes out of $f_1 + f_2$ trials, with Bonferroni-adjusted significance level $\alpha = 0.05 / m$.
3. Convert the binomial proportions to $[\mathrm{LRC}_-, \mathrm{LRC}_+] = [\log_2 (n_2\, \pi_- / n_1\, (1 - \pi_-)), \log_2 (n_2\, \pi_+ / n_1\, (1 - \pi_+))]$.

4. If the test is not significant ($LRC_- \leq 0 \leq LRC_+$), set LRC = 0. Otherwise, set LRC = LRC_- if $p_1 > p_2$ and LRC = LRC_+ if $p_1 < p_2$.

LRC has several **advantages** over other keyness measures: (i) it balances out the high-frequency bias of hypothesis tests and the low-frequency bias of effect-size measures (cf. right panel of Fig. 2); (ii) unlike heuristics such as SimpleMaths it does this in a mathematically well-justified way; (iii) it can be applied to candidates with $f_2 = 0$ without special precautions; (iv) it detects both positive ($p_1 > p_2$) and negative ($p_1 < p_2$) keywords; (v) it includes a reliable significance filter (LRC = 0) and does not require arbitrary frequency thresholds; (vi) robust and efficient implementations of the underlying binomial confidence intervals are available in standard statistical software packages, so very large candidate sets can easily be processed. The left panel of Fig. 2 shows that LRC overlaps well with established keyness measures, again indicating that it provides an excellent compromise.

A reference implementation of LRC is available at https://osf.io/cy6mw/ together with a more detailed analysis. It is also included in version 0.6 of the *corpora* package for R. [4]



Quantitative analysis of top-250 keyword lists for the data of Evert et al. (2018): overlap between four measures (left panel) and frequency distribution in the target corpus (right panel).

Bibliography

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum Books.

Culpeper, J. (2009). Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. *International Journal of Corpus Linguistics*, **14**(1): 29–59.

Dunning, T. E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1): 61–74.

Egbert, J. and Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora*, **14**(1): 77–104.

Evert, S. (2004). *The statistics of word cooccurrences: Word pairs and collocations*. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, doi: 10.18419/opus-2556.

Evert, S., Dykes, N. and Peters, J. (2018). A quantitative evaluation of keyword measures for corpus-based discourse analysis. Presentation at the Corpora & Discourse International Conference (CAD 2018), Lancaster, UK, https://www.stephanie-evert.de/PUB/EvertEtc2018_CAD_slides.pdf.

Fay, M. P. (2010). Two-sided exact tests and matching confidence intervals for discrete data. *The R Journal*, **2**(1): 53–58.

Fidler, M. and Cvrček, V. (2015). A data-driven analysis of reader viewpoints: Reconstructing the historical reader using keyword analysis. *Journal of Slavic Linguistics*, **23**(3): 197–239.

Gabrielatos, C. and Marchi, A. (2012). Keyness: Appropriate metrics and practical issues Presentation at the Corpora and Discourse Studies Conference (CADS 2012), Bologna, Italy, https://www.researchgate.net/publication/261708842_Keyness_Appropriate_metrics_and_practical_issues.

Gries, S. Th. (2005). Null-hypothesis significance testing of word frequencies: A follow-up on Kilgarriff. *Corpus Linguistics and Linguistic Theory*, **1**(2): 277–94.

Hardie, A. (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, **17**(3): 380–409.

Hardie, A. (2014). A single statistical technique for keywords, lockwords, and collocations.

Kilgarriff, A. (2009). Simple maths for keywords. *Proceedings of the Corpus Linguistics 2009 Conference*. Liverpool, UK, http://ucrel.lancs.ac.uk/publications/CL2009/.

McEnery, T., McGlashan, M. and Love, R. (2015). Press and social media reaction to ideologically inspired murder: The case of Lee Rigby. *Discourse and Communication*, **9**(2): 1–23, doi: 10.1177/1750481314568545.

Oakes, M. P. and Farrow, M. (2006). Use of the chi-squared test to examine vocabulary differences in English language corpora representing seven different countries. *Literary and Linguistic Computing*, **22**(1): 85–99, doi: 10.1093/llc/fql044.

Paquot, M. and Bestgen, Y. (2009). Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. In Jucker, A., Schreier, D. and Hundt, M. (eds), *Corpora: Pragmatics and Discourse. Papers from the 29th International Conference on English*

Language Research on Computerized Corpora. Amsterdam: Rodopi, pp. 247–69, doi: 10.1163/9789042029101_014.

Pojanapunya, P. and Watson Todd, R. (2018). Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*, **14**(1): 133–67.

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. *Proceedings of the ACL Workshop on Comparing Corpora*. Hong Kong, pp. 1–6.

Scott, M. (1997). PC analysis of key words – and key key words. *System*, **25**(2): 233–45.

Notes

1. https://www.lexically.net/wordsmith/
2. https://www.laurenceanthony.net/software/antconc/
3. https://www.sketchengine.eu/
4. https://cran.r-project.org/web/packages/corpora/

Historical Research meets Semantic Interoperability: The Documentation System SYNTHESIS and its Application in Art History Research

Fafalios, Pavlos
fafalios@ics.forth.gr
Centre for Cultural Informatics, Institute of Computer Science, FORTH, Greece

Introduction and motivation

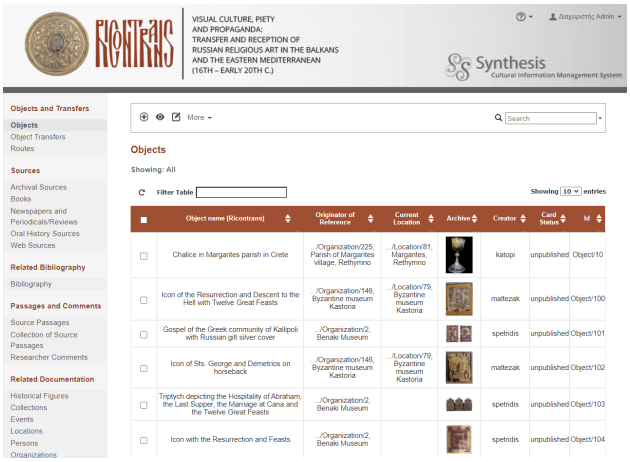
Historical science is the field that describes, examines and questions a sequence of past events, and investigates patterns of cause and effect. Research in the field usually starts by first discovering, collecting, documenting and organizing historical sources, such as written documents or material artifacts. This often includes either the transcription (and then curation) of historical archival sources, like in Petrakis et al. (2020) for the case of Maritime History, or the detailed documentation of cultural artifacts and related evidence, like in Fafalios et al. (2021) for the case of Art History, with the latter being the focus of this presentation.

In this context, although computing in the field has developed enormously over the last years, data management problems still exist and are very varied. Common problems include: a) the difficulty for collaborative but controlled documentation by a large number of historians of different research groups; b) the lack of representation of the details from which the documented relations are inferred, important for the long-term validity of the research results; c) the difficulty to combine and integrate information extracted from multiple and diverse information sources; d) the difficulty of third parties to understand and re-use the documented data, resulting in the production of data with limited longevity.

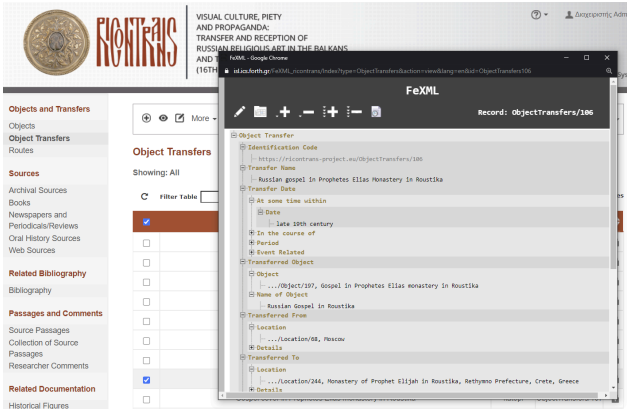
The SYNTHESIS system

In an effort to cope with the aforementioned problems, we present the SYNTHESIS documentation system and its use by a large number of historians in the context of a large European research project of Art History, called RICONTRANS (ERC Consolidator Grant, No 818791). SYNTHESIS is Web-based, multilingual, configurable (for use in other digital humanities fields), and utilizes XML technology, offering flexibility in terms of versioning, workflow management and data model extension. It focuses on semantic interoperability (Ouksel and Sheth, 1999), enabling the exchange of data among computer systems with unambiguous/shared meaning, and achieves this by making use of standards for data modelling and publication, in particular the formal ontology CIDOC-CRM (ISO 21127:2014) and the data model RDF (W3C Recommendation). The aim is the production of data with high value, longevity and long-term validity that can be (re)used beyond a particular research activity.

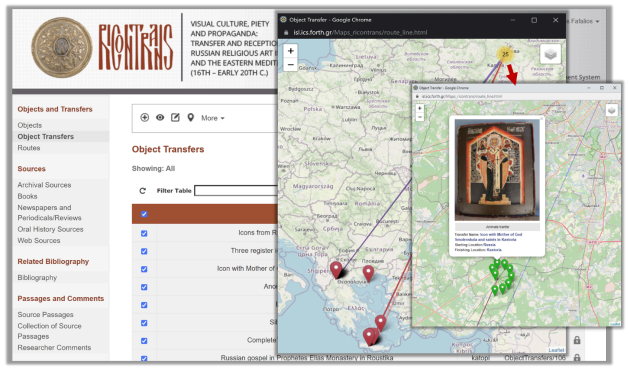
SYNTHESIS offers a wide range of functionalities including i) interlinking of the documented entities (forming a network of interrelated entities), ii) management of static and dynamic vocabularies, iii) linking to thesauri of terms, iv) connection with geolocation services (TGN, Geonames), v) map visualization for certain types of entities, vi) support of comparable historical time expressions (e.g., ca. 1920, early 16th century), vii) management of digital files (images, etc.), viii) transformation of the documented information to a knowledge base of Linked Data (Heath and Bizer, 2011).



The user interface of SYNTHESIS displaying the supported entity types (on the left) and the table of documented entities of type ‘Object’ (on the right)



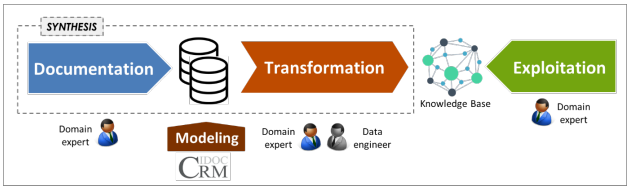
Viewing the documentation card of an entity of type ‘Object Transfer’ in SYNTHESIS



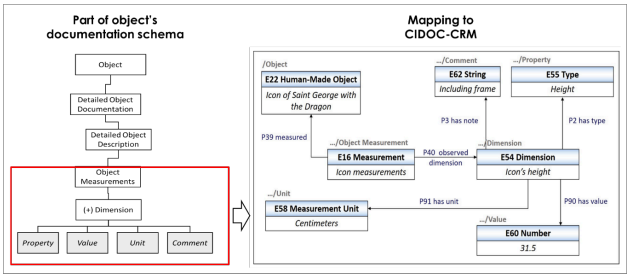
Visualizing a set of object transfers in a map

Overall philosophy and novelty

SYNTHESIS provides full-fledged support for the complete knowledge production life cycle in historical research, enabling the inclusion of rich provenance information (metadata) about the documented data and providing embedded processes for transforming the data to a Knowledge Base (KB) of historical information that is compliant to CIDOC-CRM. Contrary, to approaches that support users in creating a KB from the beginning, such as ResearchSpace (Oldman and Tanase, 2018), SYNTHESIS considers a workflow that decouples data entry from the ontology-based integration and the creation of the KB. The main reasons behind our approach are two: (a) we regard as very different a KB of facts believed together as true, versus managing and consolidating the knowledge acquisition process of a large research team. The latter requires a document structure, for making local versioning, workflow management and provenance tracing easy; (b) we consider a KB as an ideal tool for integrating the ‘latest stage of knowledge’; individual contributions, alternatives, corrections, etc., all in the same pool of valid knowledge, can hardly be regarded as a standard procedure.



The SYNTHESIS workflow: documentation and transformation to a knowledge base of historical information for exploitation in historical research



Mapping a part of object’s documentation schema to CIDOC-CRM

Usage in historical research

The system is currently used by around 40 users in five countries (mainly historians) in the context of the ongoing project RICONTRANS, whose research focus is the Russian religious artefacts brought from Russia to the Balkans after the 16th century and which are now preserved in churches, monasteries or museums. The system supports the documentation of entities belonging to totally 18 entity types, each one having its own documentation schema (data

entry fields organized in a hierarchical, tree-like structure). Indicative entity types include: objects (like icons), object transfers (like donations), archival sources, oral history sources, source passages (like a paragraph in a newspaper that talks about a topic of interest, e.g., an icon donation), historical figures (like archbishops). Currently, more than 5,000 entities have been already documented and are used in historical research, including more than 1,700 objects, 550 object transfers, 200 archival sources, 850 source passages, and 230 historical figures. By exploiting the rich connections among the documented entities, historians can find answers to complex information needs, such as "finding the routes of icons transferred to Mount Athos before the 18th century as well as the purpose of these transfers".

Lessons learned and future work

Finding the best trade-off between documentation richness and usability was a challenging problem that required extensive discussions between historians and data engineers, as well as many revisions and updates of the entity documentation schemas. On the one hand, researchers need to document information in a detailed and precise way, but on the other hand, this must be done in a quick and straightforward way. In addition, controlling the dynamic vocabularies, which allow the on-the-fly inclusion of new terms, is difficult when there is large number of editors. The problem is that we can end up with vocabularies containing multiple terms that refer to the same notion. Curation is then needed which though is laborious. A question for future work is how the system could support a better management of the dynamic vocabularies.

Acknowledgements

The following members of FORTH have been contributed in the design and development of the presented system: Konstantina Konsolaki, Lida Charami, Kostas Petrakis, Manos Paterakis, Dimitris Angelakis, Chrysoula Bekiari, Pavlos Fafalios, Martin Doerr.

The presented work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 890861 (ReKnow) and the European Research Council grant agreement No 818791 (RICONTRANS).

Bibliography

Fafalios, P., Konsolaki, K., Charami, L., Petrakis, K., Paterakis, M., Angelakis, D., Tzitzikas, Y., Bekiari, C., and Doerr, M. (2021). Towards Semantic Interoperability in Historical Research: Documenting Research Data and Knowledge with Synthesis. *Proceedings of the 20th International Semantic Web Conference (ISWC 2021)*, pp. 682-698, Springer, Cham.

Heath, T., and Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1), pp. 1-136.

Oldman, D., and Tanase, D. (2018). Reshaping the knowledge graph by connecting researchers, data and practices in ResearchSpace. *Proceedings of the 17th International Semantic Web Conference (ISWC 2018)*, pp. 325-340, Springer, Cham.

Ouksel, A. M., and Sheth, A. (1999). Semantic interoperability in global information systems. *ACM Sigmod Record*, 28(1), pp. 5-12.

Petrakis, K., Samaritakis, G., Kalesios, T., Domingo, E. G., Delis, A., Tzitzikas, Y., Doerr, M., and Fafalios, P. (2020). Digitizing, Curating and Visualizing Archival Sources of Maritime History: the case of ship logbooks of the nineteenth and twentieth centuries. *Drassana: revista del Museu Marítim*, (28), pp. 60-87.

Using Digital Tools to Detect Cross-Language Allusions in Voltaire

Gawley, James
james.gawley@gmail.com
Sorbonne University, France

This paper presents a webtool called Tesserae-OBVIL, which successfully detects cross-language allusions between the epic *Henriade* of Voltaire, written in French, and the *Aeneid* of Vergil, in Latin.

In 1717 C.E., during his imprisonment in the bastille, Voltaire began composing *La Henriade*, drawing heavily on Vergil's *Aeneid*. Voltaire's *Henriade* was designed to establish France as the cultural heir to Ancient Rome and himself as its most important writer. The poem enjoyed massive contemporary success, yet *La Henriade* receives little critical or popular attention. The problem is that Voltaire relied on a French reader with knowledge of the classical tradition and an eye for spotting allusions to classical poetry in particular. Today that reader is largely gone. In fact, the only scholar to approach the question of Voltaire's use of Vergil is O.R. Taylor, who included references to Vergil as footnotes to his edition (Taylor, 1970: pp.149-159). The present study uses digital tools to help recreate the expertise that Voltaire took for granted.

Tesserae-OBVIL is an online search tool designed to search for Latin allusions in French poetry. Unlike other tools for similarity detection and cross-language document alignment, Tesserae-OBVIL does not look for passages with the largest proportion of shared language. Because poetic allusion is intentionally subtle, it is necessary instead to search for a small number of words that are meaningful indicators of a connection. In its most common search pattern, Tesserae-OBVIL locates places where two or more words are shared between a target and source poem, within a single line or phrase. Meaningful allusions are sorted from coincidental language by filtering the search results according to the rarity of the shared words and their proximity to one another within the line or phrase.

Tesserae-OBVIL was derived from an open-source project called Tesserae, developed at the University at Buffalo (http://tesserae.caset.buffalo.edu/). When it was first deployed, the original Tesserae software was considered successful because it was able to reproduce roughly 32% of the known parallels between Lucan's *Pharsalia* and Virgil's *Aeneid* (Coffee, 2018) . Tesserae search results were then used to substantively expand the list of known allusions between these poems. Tesserae's cross-language tool fared more poorly, recovering some 11% of Latin-to-Greek allusions between Virgil and Homer (Gawley, Forstall, and Clark, 2014).

The Tesserae-OBVIL software is much more successful at detecting cross-language allusion than the original Tesserae software. To perform a search, words in French are matched with 'translations' in Latin, which are essentially cross-language synonyms. The basic search algorithm is unchanged: for a match to exist, there must be two words shared between the poems within a single phrase; for the match to score well, those words must be relatively rare and they must be close together.

The Tesserae-OBVIL search results were compared against a list of allusions between Voltaire and Virgil assembled by O.R. Taylor (Taylor, 1970). Tesserae-OBVIL was able to recover 69% of Taylor's list. The full set of search results can be retrieved at https://obvil.huma-num.fr/tesserae-obvil/cgi-bin/read_bin.pl?session=0000004a. The missing 31% largely comprises parallels where shared language is more figurative than direct. For example, in line 160 of the Henriade I, the French word 'matelots' is not quite a translation of the Latin *manus iuventum* ('band of youths'), nor is 's'empressent' really a synonym for *emicat* ('leapt'), though in both cases the same underlying concept is being described: "Les matelots ardents s'empressent sur le bord ;" *Hen*. I, 160. Cf. *Iuvenum manus emicat ardens / litus in Hesperium;* ("The band of youths leapt eagerly to the Hesperian shore;") *Aen*. VI, 5-6. Yet even with these exceptions, the success rate of the Tesserae-OBVIL software is still more than twice that of the original Tesserae software in a Latin-to-Latin context, and roughly six times as successful as Tesserae's cross-language module.

The allusions discovered using Tesserae-OBVIL add a great deal of meaning to our reading of *La Henriade*. Voltaire's biting irony is one of the reasons for his enduring popularity, but because in this case it comes in the form of complex allusions, the poet's trademark wit has hitherto passed unnoted in the scholarly literature. Some explicitly deny the presence of irony in the text (Lahouati and Mironneau, 2002: pp. 4). Yet as this presentation will demonstrate, irony is very present in Voltaire's first epic poem. Voltaire inverts of classical tropes and recasts heroic roles in a way that makes an ironic statement about France as an empire and about the reliability of authority. Both classicism and religion were important sources of authority in the world in which he composed the *Henriade*; Voltaire repeatedly references those authorities in contexts that deflate their importance.

La Henriade was conceived and composed in the Bastille, where Voltaire was imprisoned for writing verses critical of the regent. It is interesting to consider the layered irony of Voltaire's allusions in light of this imprisonment. Perhaps his sudden, indefinite confinement unnerved Voltaire, so that he preferred ambiguous praise to direct criticism at this moment in his life. Or perhaps his patriotic aim in composing *La Henriade* was quite sincere, yet complicated by anxieties over the absolute power of a monarch. Whatever his motives, Voltaire's use of classical literature adds an ironic undertone to the *Henriade*, and that ironic voice may account for the poem's tremendous success among an audience who appreciated its allusions. The Tesserae-OBVIL software presented here offers us the opportunity to recover that ironic context and, we may hope, to restore some of *La Henriade*'s notoriety.

Bibliography

Coffee, N. (2018). "An Agenda for the Study of Intertextuality", *Transactions of the American Philological Association*, 148.1: 205-223.

Gawley, J., Forstall, C., and Clark, K. (2014). "Automating the Search for Cross-Language Text Reuse." *Digital Humanities*. Lausannem, July 2014.

Lahouati, G. and Mironneau, P. (2002). "De l'urgence de lire *La Henriade*", *Voltaire Revue*, 2002.

Taylor, O.R. (1988). *La Henriade*, Genève.

Reforming the 'Eng Lit' canon: Measuring the myths and realities of English literary studies in India through a computational analysis of university curricula.

Ghosh, Arjun
arjunghosh@hss.iitd.ac.in
Indian Institute of Technology Delhi, India

The study of English Literature famously commenced not in the metropolis but in the colony – as a result of the Maculayan policies for training of would be Indian civil servants in "what is best worth knowing" through an insight into "English .. tastes, ... opinions, ... morals and ... intellect" (Macaulay, 1835; Viswanathan, 2014). Though intended for the civil servants the policy of an Anglophonic education for India created a hegemony of a western educated middle class that while being subordinated to their European rulers considered themselves superior to the 'native' underclass (Bhabha, 1994). The conjoined hegemony of the study of English language and literature persisted post-Independence with the canonical curriculum being retained in almost all Indian universities well into the 1980s – with only texts by British authors – mostly male – being included on the syllabi. The prestige and the opportunities that surrounded English education worked to maintain the preserve of an elite class who controlled the access to India's top institutions. The rise of postcolonial scholarship gradually altered the syllabi followed by department of English in India – at venturing out of the 'Brit Lit'canon to include other Literatures in English – but still from within the former colonies of Britain viz. America, Australia, Canada the Carribean and India (Rajan, 1986). 1990s onwards, a series of socio-political shifts saw the widening of the scope and reach of education among the underprivileged communities. A clamour grew among leading scholars to further democratise the curriculum by including texts from languages other than English – literature in the various Indian languages, European literature all in translation, as well as bringing in more texts written by women, dalits and black authors (Trivedi, 1995).

The current assessment undertakes a large data analysis of the current state of curricula followed by various departments of English in India institutions. The texts were classified according to their dates of publication, the gender and country of belonging of the author and the course titles. In addition to this we also mapped the examination papers for the National Eligibility Test for Lecturership (NET) which is mandatory for all university faculty, the English paper for Union Public Service Commission (UPSC) which is the entrance requirement to the Indian Administrative Services, and doctoral thesis awarded by the chosen universities as available on the ShodhGanga platform. The findings show that the movement away from the 'BritLit' canon is not as wide spread as would be believed in the leading scholarship in postcolonial studies (Rajan, 2008; Dutta, 2018), with peripheral institutions still adhering to the canon. Earlier studies that closely looked at syllabi of select universities have focussed on the changes in the syllabi and have failed to emphasize the continued adherence to the canon.

We accessed the PDF files or scans of the syllabi and converted into data tables. Scanned files were converted to digital text using OCR. Further, multiplicity of spellings of authors and texts were standardized and matched to identification data like authro's date of birth, place or year of publication etc. through a SPARQL query on wikidata. Our study through a large scale statistical and network analysis is able to put the changes in perspective and highlight the unfinished agenda of decolonization and deracialization of literary studies in India. While a handful of universities have diversified the curriculum, even in such universities areas like literature by dalits, women, queer or disabled authors are predominately included in elective courses with core courses displaying greater adherence to the canon. We also find that the spectrum of deviation from the canon is at its lowest end with the NET and UPSC examinations. Given that the NET and UPSC examinations are largely bureaucrat driven instead of being guided by the latest academic research and pedagogy, this points towards a sorry state of academic freedoms in India today. Thus what began as a curriculum designed to establish a racially defined administrative service, continues even today after seven decades of Independence.

Bibliography

Bhabha, Homi K. (1994). The Location of Culture. Routledge.

Dutta, Nandana. (2018). "View from Here – English in India: The Rise of Dalit and NE Literature." English: Journal of the English Association 67 (258): 201–8. https://doi.org/10.1093/english/efy025.

Macaulay, Thomas Babington. (1835). "Minute on Education." 1835. http://www.columbia.edu/itc/mealac/pritchett/00generallinks/macaulay/txt_minute_education_1835.html.

Rajan, Rajeswari Sunder. (1986). "After 'Orientalism': Colonialism and English Literary Studies in India." Social Scientist 14 (7): 23–35. https://doi.org/10.2307/3517248.

———. (2008). “English Literary Studies, Women’s Studies and Feminism in India.” Economic and Political Weekly 43 (43): 66–71.

Trivedi, Harish. (1995). Colonial Transactions: English Literature and India. Manchester University Press.

Viswanathan, Gauri. (2014). Masks of Conquest: Literary Study and British Rule in India, Twenty-Fifth Anniversary Edition. Columbia University Press.

An Adaptive Methodology: Machine Learning and Literary Adaptation

Glass, Grant
grantg@live.unc.edu
University of North Carolina at Chapel Hill, United States of America

Introduction

Using one of the most adapted texts in history, *Robinson Crusoe,* I ask whether or not a computer can find adaptations that scholars have yet to identify. Through testing the effectiveness of different machine learning techniques for text embedding on small groups of full-length texts, I determine the best model for our task, the universal sentence encoder, and then use it to build a deep neural network based binary classifier trained on a large dataset of adaptation and random texts. I attempt to *implicitly* teach the computer the plot of Crusoe, instead of making decisions based on stylistic details, as is a pitfall of traditional techniques. It is my hope that this novel pipeline will help other scholars work with large units of text at the plot level. Works like, Daniel Shore’s *Cyberformalism: Histories of Linguistic Forms in the Digital Archive,* Andrew Piper’s. *Enumerations: data and literary study* , and Ted Underwood’s *Distant horizons: digital evidence and literary change* all have attempted to change the literary methodology by using algorithms to find patterns and features in texts. While these methodologies utilize many machine learning techniques, these methods have met with massive pushback from the larger humanities community.[1] At the same time, these new methodologies force scholars to think about modeling and conceiving of literary texts differently (McCarty). In this shift of modeling literary text, the question that often comes up is how we can frame these literary questions in a format that a machine learning algorithm can understand.

This problem might be best considered against Daniel Defoe’s *The Life and Surprising Adventures of Robinson Crusoe*, which has never been out of print in its over three-hundred year print history and has amassed thousands of editions – not to mention the plethora of movies and T.V. shows. Using this text allows us to gain enough material to make these machine learning algorithms viable. Teaching a machine learning algorithm the story of Crusoe (by feeding it different adaptations) we could ask if it could start to distinguish between a random story and a Crusoe like story? Could it identify new adaptations of Crusoe that have yet to be discovered?

Data Description

In early experiments to determine the suitable text embedding technique, I used four different texts: the “Original Text” which is the 1719 first edition of *Robinson Crusoe* by Daniel Defoe, a “close” adaptation[2] of Jenichiro Oyabe’s *A Japanese Robinson Crusoe,* a “far” science fiction adaptation[3] called *The Happy Castaway* (1965)*,* and a random text, *Pride and Prejudice* by Jane Austen (1813). I chose the text based on my own scholarship of Robinson Crusoe: the “close” adaptation is one that follows the exact storyline, but reimagines the story as a Japanese man in America, the far adaptation takes the same plot, but everything about the text is changed, and the random text is something similar stylistically, but has no character or plot similarity to Crusoe.

The final project utilized two different large datasets, the first corpus of which was a random pooling of 2,188 texts from the Eighteenth Century Collections Online (ECCO) Text Creation Partnership (TCP)[4]. This data is freely available through the ECCO-TCP website and was verified through the corresponding CSV file, a preview of which is included in Table 1.

TCP	EEBO	VID	STC	Status	Author	Date	Title
K000102.000				Free	Centlivre, Susanna, 1667?-1723.	1705	The basset-table: A comedy. As it is acted at the Theater Royal in Drury Lane, by Her Majesty's servants. By the author of The gamester.
K000103.000				Free	Jones, Henry, 1721-1770.	1745	The bricklayer's poem to the Countess of Chesterfield, on Her Ladyship's saving the soldiers from being shot
K000102.000				Free	Congreve, William, 1670-1729.	1705	Prologue to the court: on the Queen's birthday. 1704.
K000100.000				Free	Cumberland, Richard, 1732-1811.	1772	The fashionable lover: a comedy: as it is acted at the Theater Royal in Drury Lane.
K000266.000				Free	Defoe, Daniel, 1661?-1731.; Pitts, William, 1674-1724.	1705	The diet of Poland: a satyr. Considered paragraph by paragraph. To which is added a key to the whole, ...
K000268.000				Free	Duck, Stephen, 1705-1756.	1141	Every man in his own way: An epistle to a friend. By Stephen Duck.
K000269.000				Free	Gedoyn, William, 1777-1797.	1749	An essay upon rural and the management of children. Taken from birth to three years of age. By a physician. In a letter to one of the governors of the Foundling Hospital. Published by order of the General Committee ...
K000342.000				Free	Alderson, John, 1757-1829.	1794	An essay on the thus toxicodendron: pubescent poison oak, or sumach, with cases showing it is efficacy in the cure of paralysis, ... By John Alderson, ...
K000179.000				Free	Duck, Stephen, 1705-1756.	1141	Hints to a schoolmaster. Addressed to the Revd. Dr. Turnbull. By Stephen Duck.
K000400.000				Free	Hill, John, 1714?-1775.	1775	Hypochondriada. A practical treatise on the nature and cure of that disorder, commonly called the hyp and hypo. By Sir John Hill
K000413.000				Free	Kelly, Hugh,	1760	An elegy to the memory of the Right Honorable William, late Earl of Bath
K000444.000				Free	Carleton, George, fl. 1728.; Defoe, Daniel, attributed name. 1661?-1731.	1728	The memoirs of an English officer: who served in the Dutch war in 1672. to the peace of Utrecht, in 1713. ... By Capt. George Carleton.
K000222.000				Free	Parnell, Thomas, 1679-1718.	1713	The hymn and the olive: or, war and peace.
K000187.000				Free	Aikenhead, John, 1661-1736.	1712	John Bull still in his senses: being the third part of Law is a bottomless pit. Printed from a manuscript found in the cabinet of the famous Sir Humphry Polesworth: and published, (as well as the two former parts) by the author
K000063.000				Free	Centlivre, Susanna, 1667?-1723.	1713	An epistle to Mrs. Wishup: now in the train of Her Royal Highness, the Princess of Wales. As it was sent to her to the Hague. Written by Mrs. Susanna Centlivre.
K000065.000				Free	Hill, Aaron, 1685-1750.	1707	Camilla: a poem: humbly inscribed to the Right Honorable Charles Earl of Peterborough and Monmouth. By Aaron Hill, gent.
K000061.000				Free	Fielding, Henry, 1707-1754.	1730	The coffee-house politician: or, the justice caught in his own trap. A comedy, As it is acted at the Theater Royal in Lincoln's Inn Fields. Written by Mr. Fielding.; Rape upon rape
K000760.000				Free	Cumberland, Richard, 1732-1811.	1764	An elegy written on Saint Mark's eve
K000781.000				Free	Gentleman, Francis, 1728-1784.	1769	The general: A poem: Respectfully inscribed to the Right Honorable the Marquis of Granby. By the author of A trip to the moon.
K000411.000				Free	Fordyce, George, 1736-1802.	1772	Elements of the practice of physic: Part the first. Containing the natural history of the human body. By George Fordyce, ... Elements of the practice of physic.
K000113.000				Free	Savage, Richard, d. 1743.	1732	An epistle to the Right Honorable Sir Robert Walpole, ... By Richard Savage esquire.
K000023.000				Free	Maley, John, d. 1776.	1763	An essay on the nature and cure of the king's evil: deduced from observations and practice. The second edition: with an addition of remarkable cases of poor sufferers, cured by the author. ...
K000025.000				Free	Hope, Alexander, 1889-1916.	1739	The history of the Norfolk steward contin.ued: in two parts. Part I. Containing an account of Mr. Lyn's private character, ... Part II. Containing some farther account of Mr. Lyn's management, ...
K000054.000				Free	Pain, Thomas, 1737-1809.	1796	The American crisis, and a letter to Sir Guy Carleton, on the murder of Captain Huddy, and the intended retaliation on Captain Asgill, of the Guards. By Thomas Pain,
K000044.000				Free	Churchill, Charles, 1731-1764.	1762	An epistle to the author of The four letters to Churchill. By the author of The Revolt of C--n-G-rd-n.
K000059.000				Free	Hamble, John Philip, 1751-1823.; Bradwell, Thomas, 1642?-1692.	1763	The female officer: or the humors of the army, a comedy: Altered from Bradwell.
K001031.000				Free	Pott, Percivall, 1714-1788.	1765	An account of a particular kind of rupture, frequently attended upon newborn children: and sometimes met with in adults: viz. that in which the intestine, ... is found in the same cavity, and in contact with the testicle. By Perci
K001006.000				Free	Hoken, William, 18th cent.	1788	An essay on humanity: or a view of abuses in hospitals. With a plan for correcting them. By William Hoken.
K001048.000				Free	Defoe, Daniel, 1661?-1731.	1713	An answer to The tale of a nettle: Written by D. D'Foe.
K001000.000				Free		1772	An essay on the forms of imagination in pregnant women: Addressed to the ladies.
K001153.000				Free	More, Hannah, 1745-1833.	1795	The carpenter: or, the danger of evil company.
K001176.000				Free	Moultrie, Arthur, 1727-1805.	1769	Genuine memoirs of the life and adventures of the celebrated Miss Ann Elliot. Written by a gentleman intimately acquainted with her. ...
K001004.000				Free	Francis, Philip, Sir, 1740-1818.	1787	House of Commons, Tuesday, 11th December, 1787. Mr. Francis. Mr. Speaker, before I offer any thing to the consideration of the House
K001007.000				Free	Cheyne, George, 1673-1743.	1721	An essay on the gout: with an account of the nature and qualities of the Bath waters. Intended for the benefit of Richard Tennison, Esquire; By Geo. Cheyne, ... Observations concerning the nature and due method of treating
K001009.000				Free	Defoe, Daniel, 1661?-1731.	1705	Advice to all parties: By the author of The true-born Englishman.
K001012.000				Free	O'Keeffe, John, 1747-1833.	1783	The agreeable surprise: A comic opera, In two acts. By Mr. O'Keeffe. The music composed by Dr. Arnold.
K001024.001				Free	Goldsmith, Oliver, 1730?-1774.	1762	The citizen of the world: or, letters from a Chinese philosopher, residing in London, to his friends in the east. ...
K001024.002				Free	Goldsmith, Oliver, 1730?-1774.	1762	The citizen of the world: or, letters from a Chinese philosopher, residing in London, to his friends in the east. ...
K001060.000				Free	Bertanzoni, John, 1730?-1791.	1783	An essay on the bite of a mad dog: in which the claim to infallibility of the principal preservative remedies against the hydrophobia is examined. By John Berkenhout, M.D.
K001062.000				Free	Jephson, Robert, 1736-1803.	1783	The hotel: or, the servant with two masters. As it was performed at the Theater Royal, Smock-Alley, with distinguished applause
K001115.000				Free	Hill, Aaron, 1685-1756.	1709	The invasion: a poem to the Queen. By Mr. Hill.

Table 1:
ECCO-TCP CSV File describing all the data in the corpus.

The next corpus included 1,484 texts drawn from a variety of variations of *Robinson Crusoe,* pulled from HathiTrust[5] using Hathitrust's Rsync.

Core Methodology

The first experiment to determine which method would generate the best text embeddings for this task was with the sklearn TfidfVectorizer to build the embeddings of our training data[6]. I calculated the cosine similarity scores for each of the texts to the reference text (i.e. the Original Text) using sklearn's metrics-pairwise package. Then I experimented with Google Research's Universal Sentence Encoder (USE)[7]. While originally meant for generation of sentence-level embeddings, the model does not actually require a set maximum sequence length, which is a useful functionality that allows us to represent full-length texts of varying lengths as a fixed-dimensional embedding layer. In the end, I chose to work with the USE embeddings because it gave more context-aware, and as a result, discriminatory embeddings than the other candidates. Notice (in Figure 1) that the text determined as 'close' to the reference text by me (human expert), while indeed the closest, still showed a cosine similarity of only 0.528. Further, the texts determined as 'random' and 'far' were also significantly further from 'close' as well as the reference text, but very close to each other - which is what we might expect from a model which has learnt semantic relationships particularly well (after all, why should *Pride and Prejudice* be closer to *Robinson Crusoe* than *The Happy Castaway* - both are unrelated by plot). Note that the BERT embeddings, *Pride and Prejudice* turned out to be closer to *Robinson Crusoe* which we posit is due to the nature of the sentence-level embeddings - the representation learnt is more about the similarity in the stylistic/linguistic/grammatical/lexical sense than about the plot.



Figure 1:
Results of Initial Method Across Different Texts (1.0 being closest to the original text) USE-Universal Sentence Encoder.

Final Results

The model performs exceptionally well on the validation and test sets, identifying the adaptations (denoted by class 1 in Figure 2) with near perfect precision and recall.

	precision	recall	f1-score	support
0	0.98	1.00	0.99	137
1	1.00	0.99	1.00	309
accuracy			0.99	446
macro avg	0.99	1.00	0.99	446
weighted avg	0.99	0.99	0.99	446

Figure 2:
The classification results from the mode. 0 denotes non-adaptations and 1 denotes adaptations

Current Conclusions and Future Work

The potential pitfall with this technique is that I will not be able to measure how similar a text is to *Robinson Crusoe*, on a more textual level, but I have already attempted this in previous work using doc2vec. By using this new technique to look at a larger window of text than a sentence, we can find works that share a similar plot, which would begin to make a new model of adaptation centered around plot rather than setting or characters. The challenge becomes where exactly the plot gets figured out, what unit of text can tell us that? If we can begin to think about where the plot gets encoded in the text and we can make the window of analysis the same, then we can begin to move forward to other machine learning problems.

Bibliography

Chaudhary, Vishrav, et al. "Low-Resource Corpus Filtering using Multilingual Sentence Embeddings." *arXiv preprint arXiv:1906.08885* (2019).

Christenson, Heather. "HathiTrust." *Library Resources & Technical Services* 55.2 (2011): 93-102

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

Foster, Thomas C., David De Vries, and © 2012 by Harper Collins Publishers. *How to read literature like a professor*. Harper Collins Publishers, 2012.

Frye, Northrop. *Anatomy of criticism: Four essays*. Princeton University Press, 2020.

McCarty, Willard. *Humanities computing*. 2014.

Moretti, Franco. *Distant reading*. Verso Books, 2013.

Piper, Andrew. *Enumerations: data and literary study*. University of Chicago Press, 2018.

Rae, Jack W., et al. "Compressive transformers for long-range sequence modelling." *arXiv preprint arXiv:1911.05507* (2019).

Sanders, Julie. "Adaptation/Appropriation." *The Encyclopedia of the Novel* (2010).

Shore, Daniel. *Cyberformalism: Histories of Linguistic Forms in the Digital Archive*. JHU Press, 2018.

Underwood, Ted. *Distant horizons: digital evidence and literary change*. University of Chicago Press, 2019.

Watt, Ian. *The rise of the novel*. Univ of California Press, 2001.

Welzenbach, Rebecca. "Making the Most of Free, Unrestricted Texts: a first look at the promise of the Text Creation Partnership." (2011).

Yang, Yinfei, et al. "Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax." *arXiv preprint arXiv:1902.08564* (2019).

Notes

1. See Da, Nan Z. "The computational case against computational literary studies." *Critical inquiry* 45.3 (2019): 601-639.
2. “Close” refers to the similarity in setting, characters, and plot to the original.
3. “Far” refers to only the plot being loosely similar to the original text.
4. https://textcreationpartnership.org/tcp-texts/ecco-tcp-eighteenth-century-collections-online/ see Welzenbach, Rebecca. "Making the Most of Free, Unrestricted Texts: a first look at the promise of the Text Creation Partnership."
5. https://www.hathitrust.org see Christenson, Heather, "HathiTrust.”
6. https://scikit-learn.org/stable/modules/classes.html#module-sklearn.feature_extraction.text We used the S
7. https://tfhub.dev/google/universal-sentence-encoder/4- the latest pretrained model available, updated 2020. See Yang, Yinfei, et al. "Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax."

Optical Character Recognition for Complex Scripts: A Case-study in Cuneiform

Gordin, Shai
shaigo@ariel.ac.il
Ariel University, Israel

Romach, Avital
avitalromach@tauex.tau.ac.il
Tel Aviv University, Israel

Cuneiform is one of the earliest writing systems in the world, usually written by pressing a stylus on moist clay tablets, creating a three-dimensional script. It is a logo-syllabic script, like Chinese or Japanese writing systems (Veldhuis, 2012; Kwan, 2014; Francis, 2020). There are close to a thousand cuneiform signs, not all of which were used simultaneously—usually about 200-300 signs were used at once. Additionally, the forms of these signs changed drastically during the 3,000 years in which this writing system was in use. Cuneiform tablets are available for OCR in three main formats: (a) 3D or 2D+ models, the most accurate representation, but these are still relatively rare and expensive to produce; (b) 2D images of cuneiform objects, which are fast growing in quality and ubiquity; and (c) hand-copies, 2D drawings made by scholars of the inscribed signs. These are still commonly used in the field, and are the majority of available representations of cuneiform writing. Furthermore, cuneiform writing and all the languages using it, are low-resource languages, which raises difficulties when implementing some computational methods (Hedderich et al., 2021). Although the number of digital texts is steadily growing in recent years, the gap is still substantial, and significant efforts are needed for digitization.

The best results for character recognition of cuneiform signs or strokes thus far have been achieved from 3D models (Mara et al., 2010; Fisseler et al., 2013; Fisseler et al., 2014; Rothacker et al., 2015). Work on hand-copies, or 2D projections of 3D models which look like hand-copies, are also showing promising results (Mara and Krömker, 2013; Bogacz et al., 2015a; Bogacz et al., 2015b; Bogacz et al., 2016; Massa et al., 2016; Bogacz and Mara, 2018; Yamauchi et al., 2018), as well as some work on 2D images (Rusakov et al., 2019; Rusakov et al., 2020; Dencker et al., 2020; see also Bogacz and Mara, forthcoming). These results, nevertheless, are limited in scope because of lack of labeled data, lack of 3D models, or, most often, a lack of an

accessible way for cuneiform specialists to use the models developed.

With this gap in mind between the specialists and the code, the set of CuRe tools (Cuneiform Recognition) of the Babylonian Engine project was created as an online interactive platform for scholars. [1] The idea was that already in the design and initial training of machine learning (ML) models, humans, particularly experts, should take an active part. The ML models are envisioned as "co-workers" which provide likely suggestions to the user, aiding the process of cuneiform scholarly edition publication, and improving as the user corrects them. This way, it is not only the ML models that benefit from the corrections and labeled data created by experts, but also the experts can enjoy a designated work environment for any type of cuneiform text, and download the results of their work– already advancing cuneiform scholarship.

Our main tool, CuRe (Cuneiform Recognition), takes images of hand-copies as input, either uploaded by the user or one of the images in the CuRe dataset of copyright-free hand-copies of Neo- and Late Babylonian documents (see demo, the full platform will be available in the near future). Hand-copies were chosen for initial training because they are more ubiquitous than 2D images or 3D models, and many hand-copies are still missing full publication of their textual editions. The signs are detected using a Faster R-CNN ResNet 50 model (detectron2), and then given suggestions for their value with ResNet-18 (Figure 1). The two stages of the model are also split for the user. First, the user corrects the sign detection, and then they are given the suggested identifications. The models currently perform with ca. 92% accuracy for sign identifications, checked against a validated test set, and we are constantly enlarging the dataset and retraining the model.

We are in the process of developing another stage in this tool: stroke recognition from 2D images. Identifying the strokes—the constituent parts which make up a cuneiform sign—is drastically simpler than identifying the whole (similar to text recognition for Japanese in Liang et al., 2015). While the signs changed drastically in the 3,000 years in which cuneiform was in use, the strokes and writing technique remained similar. Furthermore, since there are only three main stroke types (horizontal, vertical, and oblique), it is much quicker to obtain a large corpus of examples of each, as compared to collecting enough samples for every sign and its variant forms. There have been no previous attempts of identifying strokes from 2D images. After identification using Faster R-CNN ResNet 50, vector images are created with schematic representations of the strokes, which are quite like hand-copies (Figure 2).

The final goal is for the user to create a digital text using OCR in three stages: receiving and correcting (a) stroke identification, (b) sign detection, and (c) the identification of the sign values. This process creates additional validated training data for all our models, constantly improving them, while avoiding multiplying the error rates in the final result—the digitized text. Since the models are already approximately 80%-90% accurate (and rising), the amount of corrections is not too cumbersome, and it is already able to save a significant amount of time that experts usually spend on deciphering cuneiform texts.

To conclude, designing ML models as part of a human-in-the-loop pipeline application has the following benefits: (a) people are incentivized to create the data needed for training; (b) breaking down the "reading" process of the OCR, allows for less mistakes in the final digitized text; (c) the ML models are used from their inception in a real-world scenario, providing real-world value to the research community. Furthermore, we believe the CuRe tool-set can also be a valuable learning environment for cuneiform students or any interested laypeople. The advantage for cuneiform studies is thus twofold: growing a database of available digital editions for research, and creating a learning environment which can help disseminate knowledge on some of the oldest civilizations in human history. The digital editions created will be downloadable in the standard formats of the field (ATF, TEI/XML, and JSON). We believe that this work pipeline—breaking down the OCR process and developing human-in-the-loop ML models—is an effective way for solving OCR for additional low-resource and complex writing systems (compare Hashimoto et al., 2018 on Japanese), or for that matter, NLP applications for diverse low-resource languages (Wang et al., 2021).

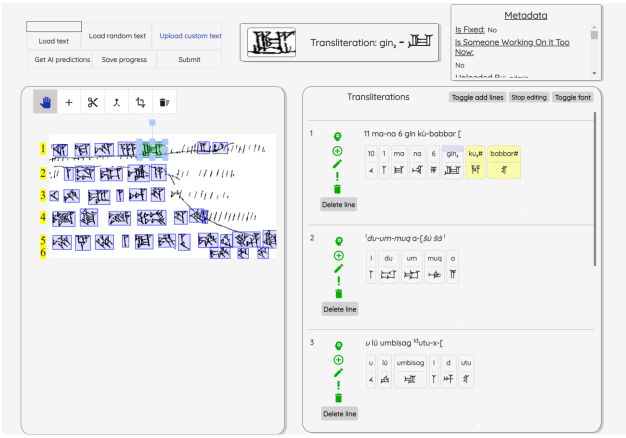


Figure 1:
A screenshot of the CuRe-tool from the Babylonian Engine website, currently in deployment, see

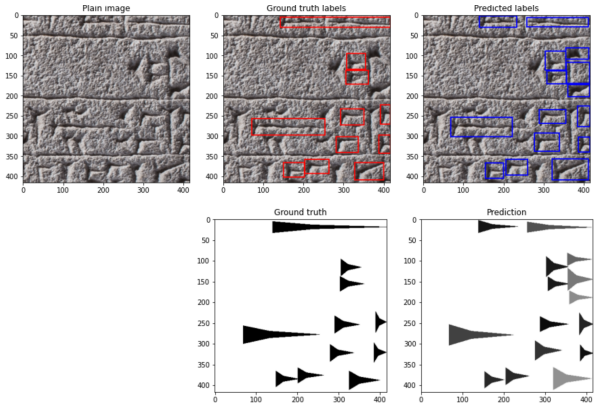


Figure 2:
A section of cuneiform writing with ground truth and identified labels for horizontal strokes (top), and a vectorized representation of the strokes (bottom). The grayscale on the strokes signifies the certainty level of the model (tablet image: © Yale Babylonian Collection, courtesy of Klaus Wagensonner)

Bibliography

Bogacz, B., Gertz, M. and Mara, H. (2015a). Character Retrieval of Vectorized Cuneiform Script. In 13th International Conference on Document Analysis and Recognition (ICDAR 2015). Piscataway, NJ: IEEE Computer Society, pp. 326–30. 10.1109/ICDAR.2015.7333777.

Bogacz, B., Gertz, M. and Mara, H. (2015b). Cuneiform Character Similarity Using Graph Representations. In Wohlhart, P. and Lepetit, V. (eds), 20th Computer Vision Winter Workshop. Graz: Verlag Der Technischen Universität Graz, pp. 105–12.

Bogacz, B., Howe, N. and Mara, H. (2016). Segmentation Free Spotting of Cuneiform Using Part Structured Models. In 15th International Conference on Frontiers in Handwriting Recognition (ICFHR 2016). Piscataway, NJ: IEEE Computer Society, pp. 301–6. 10.1109/ICFHR.2016.0064.

Bogacz, B. and Mara, H. (2018). Feature Descriptors for Spotting 3D Characters on Triangular Meshes. In 16th International Conference on Frontiers in Handwriting Recognition (ICFHR 2018). Piscataway, NJ: IEEE Computer Society, pp. 363–8.

Bogacz, B. and Mara, H. (forthcoming). Digital Assyriology - Advances in Visual Cuneiform Analysis. Journal on Computing and Cultural Heritage.

Dencker, T. et al. (2020). Deep Learning of Cuneiform Sign Detection with Weak Supervision Using Transliteration Alignment. PLOS ONE, 15(12), p. e0243039. 10.1371/journal.pone.0243039.

Fisseler, D. et al. (2013). Towards an Interactive and Automated Script Feature Analysis of 3D Scanned Cuneiform Tablets. In Scientific Computing and Cultural Heritage 2013. http://www.cuneiform.de/fileadmin/user_upload/documents/scch2013_fisseler.pdf.

Fisseler, D. et al. (2014). Extending Philological Research with Methods of 3D Computer Graphics Applied to Analysis of Cultural Heritage. In Klein, R. and Santos, P. (eds), Eurographics Workshop on Graphics and Cultural Heritage (GCH 2014). Goslar: The Eurographics Association, pp. 165–72. 10.2312/gch.20141314.

Francis, N. (2020). New Research on the Adoption and Transformation of Chinese Writing. Chinese Language and Discourse, 11(1), pp. 134–45. 10.1075/cld.19007.fra.

Hashimoto, Y. et al. (2018). Minna de Honkoku: Learning-Driven Crowdsourced Transcription of Pre-Modern Japanese Earthquake Records. In ADHO / EHD 2018 - Mexico City.https://dh-abstracts.library.cmu.edu/works/6312 (accessed 9 December 2021).

Hedderich, M. A. et al. (2021). A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. NAACL-HLT 2021. Online: Association for Computational Linguistics, pp. 2545–68. 10.18653/v1/2021.naacl-main.201.

Kwan, T.-W. (2014). Phenomenological Interpretation of the “Six Ways” of Chinese Script Formation. In Gordin, Sh. (ed.), Visualizing Knowledge and Creating Meaning in Ancient Writing Systems. Berliner Beitrage zum Vorderen Orient. Gladbeck: PeWe-Verlag, pp. 157–202.

Liang, J. et al. (2015). Character-Position-Free On-Line Handwritten Japanese Text Recognition. In 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). Kuala Lumpur, Malaysia: IEEE, pp. 231–5. 10.1109/ACPR.2015.7486500.

Mara, H. et al. (2010). GigaMesh and Gilgamesh – 3D Multiscale Integral Invariant Cuneiform Character Extraction. In Artusi, A. et al. (eds), The 11th International Symposium on Virtual Reality, Archaeology and Cultural Heritage. VAST: International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage. The Eurographics Association. 10.2312/VAST/VAST10/131-138.

Mara, H. and Krömker, S. (2013). Vectorization of 3D-Characters by Integral Invariant Filtering of High-Resolution Triangular Meshes. In Proceedings of the International Conference on Document Analysis and Recognition (ICDAR 2013). Piscataway, NJ: IEEE Computer Society, pp. 62–6. 10.1109/ICDAR.2013.21.

Massa, J. et al. (2016). Cuneiform Detection in Vectorized Raster Images. In Čehovin, L. Mandeljc, R. and

Štruc, V. (eds), 21st Computer Vision Winter Workshop. Ljubljana: Slovenian Pattern Recognition Society.

Rothacker, L. et al. (2015). Retrieving Cuneiform Structures in a Segmentation-Free Word Spotting Framework. In Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing (HIP 2015). New York, NY: Association for Computing Machinery, pp. 129–36. 10.1145/2809544.2809562.

Rusakov, E. et al. (2019). Generating Cuneiform Signs with Cycle-Consistent Adversarial Networks. In Proceedings of the 5th International Workshop on Historical Document Imaging and Processing. HIP '19. New York, NY, USA: Association for Computing Machinery, pp. 19–24. 10.1145/3352631.3352632.

Rusakov, E. et al. (2020). Towards Query-by-Expression Retrieval of Cuneiform Signs. In 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 43–8. 10.1109/ICFHR2020.2020.00019.

Veldhuis, N. (2012). Cuneiform: Changes and Developments. In Houston, S.D. (ed.), The Shape of Script: How and Why Writing Systems Change. Santa Fe, N.M.: SAR Press, pp. 3–23.

Wang, Z. J. et al. (2021). Putting Humans in the Natural Language Processing Loop: A Survey. In Blodgett, S.L. et al. (eds), Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing. Online: Association for Computational Linguistics, pp. 47–52.https://www.aclweb.org/anthology/2021.hcinlp-1.8.

Yamauchi, K., Yamamoto, H. and Mori, W. (2018). Building A Handwritten Cuneiform Character Imageset. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA), pp. 719–22. https://aclanthology.org/L18-1115.

Notes

1. Funded by the Ministry of Science & Technology, Israel, Grant 3-16464.

The many faces of theory in DH: Toward a dictionary of theoreticians mentioned in DH

Gutiérrez de la Torre, Silvia E.
silviaegt@uni-leipzig.de
Computational Humanities Group, Leipzig University

Burghardt, Manuel
burghardt@informatik.uni-leipzig.de
Computational Humanities Group, Leipzig University

Niekler, Andreas
aniekler@informatik.uni-leipzig.de
Computational Humanities Group, Leipzig University

Kleymann, Rabea
kleymann@zfl-berlin.org
Leibniz-Zentrum für Literatur- und Kulturforschung Berlin

Introduction

Theory in the Digital Humanities (DH) has been the subject of much debate. While the narrative of an "end of theory" (Anderson, 2008) is still being used, new discourses have found their way into DH scholarship. On the one hand, we observe a *laboratory turn* in DH that focuses on theory in practice (see Saklofske, 2016; Pawlicka-Deger, 2020). On the other hand, theoretical endeavors are regarded as much needed forms of criticism (see Alvarado, 2019; Drucker, 2021). Despite the relevance of these strands, a systematic investigation has yet to be conducted. In our paper, we want to add to these ongoing debates by laying the foundation for further empirical approaches[1]that investigate the occurrences of *theory*references in scholarly written communication, i.e. DH journals, abstracts, and even forums such as the *Humanist*discussion group[2]. To provide an example study, we examined a corpus of 3,737 articles from three well-known DH journals: "Computers and the Humanities (CHum)", "Literary and Linguistic Computing/Digital scholarship in the Humanities (LLC)", "Digital Humanities Quarterly (DHQ)". Two of these "foundational journals" (Sula & Hill, 2019) have a fairly long history: CHum was established in 1966, and LLC in 1986; while the third, DHQ, is one of the largest journals aimed specifically at DH research and has been used before as the basis of "digital humanities" research (see Gao et al., 2018; Luhmann & Burghardt, 2021). However, we are well aware that using the English buzzword "digital humanities" implies a self-fulfilled prophecy of anglocentric bias, since it leaves out the history of parallel DH histories in other languages and geographies (Gutiérrez De la Torre, 2020). In sum, our corpus covers a time span from 1966-2020 and has a total size of approx. 19 million tokens.

A keyword lookup shows that 1,268 articles (33.9 %) contain an instance of the search strings: *theory*/*theories*. To expand the semantic field of these strings, we have built a set of dictionaries with semantically related words (see Heuser & Le-Khac, 2012). In this article, we present work in progress for the creation of a dictionary of theoreticians that are mentioned in DH, which we consider an important prerequisite for any empirical study of the function and development of theory in DH.

Building a corpus of theoreticians mentioned in DH papers

As part of an ongoing publication (Kleymann et al., under review) on a corpus-based study of "theorytellings" in DH, we handcrafted a dictionary for the sub-field of literary theory which we based on three widely-used introductory works on the topic (Selden et al., 2006; Rivkin & Ryan, 2007; Castle, 2008). The dictionary is available via GitHub[3]and covers 13 theoretical frameworks and typical representatives of both authors and related theory strands. A look-up of all these items in the corpus of DH journals yields 1,507 occurrences in 793 out of 3,737 articles. Following these promising results, our next step was to enhance the dictionary in a systematic way.

Many of the literary theory frameworks from our handcrafted dictionary can also be found in the form of categories on Wikipedia. Categories are collections of pages to a larger theme, for instance "hermeneutics"[4]or "new criticism"[5]. They contain pages with related concepts and also persons that are commonly associated with the topic. Our assumption is that theories tend to be presented by quoting their representative authors, thus we were particularly interested in retrieving persons inside each category. Creating a dictionary of persons also has the advantage that we can further enhance it with additional information from Wikidata (name, country, occupation, gender, etc.).

First, we generated a list of theory "seeds" from our corpus which were identified by simple heuristics such as "theor* of _" (e.g. theory of information) , "noun + theory" (e.g. systems theory) "adjective + theory" (e.g. economic theory). Using these heuristics, we were able to collect a total of 3,323 theory strings from our corpus.[6]These strings were used to query Wikipedia categories via the MediaWiki API[7]which looks for the given string both in the title as well as in the description snippets of all categories[8]: 1,529 different query strings were matched against 1,266 different categories, however since many strings matched to more than one category the API returned a total of 9,772 categories.[9]To reduce the fuzziness induced by the category snippets matches, we calculated the Jaccard distance and only kept category names that had a distance score lesser than 0.6.[10]This left us with 2,411 person entries[11]and 92 categories, most of which look promising in terms of matching actual theory-frameworks (i.e. actor-network theory, chaos theory, critical theory, deconstruction, film theory, etc.). In order to check how many of the newly added theory-related persons actually appear in our corpus, we did a lookup of all the names. In order to compensate for the ambiguity of some last names (e.g. *Field*) we searched only upper-case instances.

We were able to find 217 persons (11.6% of the retrieved dictionary) cited with their full names in our corpus[12] (see Table 1). Moreover, via the Wikidata reconciliation we know 187 of them are male (86.5%), 28 female (12.9%) and one, non-binary, providing extended possibilities for diversity studies (see González et al., 2021)

position	keyword	tf	df
1	Lev Manovich	34	30
2	Christopher Marlowe	44	25
3	Jay David Bolter	30	24
4	Marshall McLuhan	32	24
5	Allen Forte	29	23
6	Seneca	43	22
7	Michael Joyce	40	20
8	Michel Foucault	22	20
9	Ezra Pound	27	19
10	George Landow	28	19
11	Bruno Latour	21	17
12	Francis Bacon	41	15
13	Walter Benjamin	18	15
14	Alan Turing	17	14
15	Jacques Derrida	17	14
16	Noam Chomsky	15	14
17	Roman Jakobson	15	14
18	Boethius	27	12
19	Northrop Frye	13	12
20	Terry Winograd	13	12

Table 1: Term and document frequency of the 20 most frequent persons with a full name match in our corpus.

By looking to every possible n-gram representation of the reference name (i.e. “Schulz von Thun”, “von Thun”, “Thun”) we got 1,216 matches that correspond to 53.5 % of humans in the dictionary.[13] However, many of the frequent last names here are highly ambiguous, either because they can be confused with common first names (Thomas, James, Paul, etc.) or because they are widely used last names (Smith, Brown, West, etc.). To make use of these last name matches, we aim to adopt an advanced disambiguation approach that is based on word embeddings (Müller, 2017) in follow-up studies.

Conclusions

This first experiment for creating theory-related dictionaries by using seed terms from a corpus of three DH journals and looking them up via the MediaWiki API delivered very promising results. Not only were we able to gather a vast number of person names that are somehow related to theory-frameworks, but we were also able to show that these person names in turn can also be found in the actual DH articles. Two main limitations must be taken into account: 1) our “seed” terms are limited by literality (“theory of” statements) and do not encompass all possible theories. 2) API results depend on the robustness of Wikipedia categories and Wikidata information. So for instance, if a female Latin American theorist was not listed within a Wikipedia category, her name will not be retrieved. Conversely, if a theorist has not been labeled as “human” in Wikidata -i.e. if the property “instance of” (P31) is not filled with the entity “human” (Q5)- this theorist will not be retrieved.

Thus, we will have to come up with some routines of manually cleaning these automatically generated dictionaries. Therefore, we would be happy to get the chance to discuss our approach with the DH community as part of DH 2022 before we dive deeper into follow-up studies. In sum, we would like to share our method and reflect together on expanding this research to other DH communities of practices across other languages and geographies.

Notes

1. Alex H. Poole (2017) observes that DH tends to discuss their identity – which we believe also includes their relation to theory – on a mostly anecdotal basis. Poole suggests that “the field would benefit from exploring itself more empirically” (p. 107). One can find an antecedent of an empirical approach in Mark Hall’s DHd 2019 contribution “DH is the Study of dead Dudes”, in which persons mentioned in DHd conferences (2016, 2017, and 2018) were manually identified from the abstracts, according to the following criteria: names are only counted if their work is the primary subject of investigation (#1), if they are presented as an exemplary example of a topic (#2) or as a sample of a dataset (#3). Our article can be seen as being complementary to Hall’s work since one limitation was precisely to exclude names that are part of the methodological approach (Hall, 2019).

2. Humanist mailing list: https://dhhumanist.org/

3. Literary theory dictionary: https://github.com/theory-in-dh/conceptual_forays/blob/main/data/theory_dictionary.md

4. Hermeneutics: https://en.wikipedia.org/wiki/Category:Hermeneutics

5. New Criticism: https://en.wikipedia.org/wiki/Category:New_Criticism

6. For a list of all queries see: https://github.com/theory-in-dh/conceptual_forays/blob/main/data/1_theoriesof_complete.csv

7. Mediawiki API: https://www.mediawiki.org/wiki/API:Main_page

8. For an example of the results for the query “theory of language” see: https://en.wikipedia.org/w/api.php?action=query&list=search&srnamespace=14&srsearch=theory%20of%20language

9. See the complete database here: https://github.com/theory-in-dh/conceptual_forays/blob/main/data/2_wikipediacategoriesfromquery.csv

10.The filtered database can be found here: https://github.com/theory-in-dh/conceptual_forays/blob/main/data/3_wikicategories_distances_filtered.csv

11. A complete list of the persons with additional metadata such as gender, country, etc. is available on GitHub: https://github.com/theory-in-dh/conceptual_forays/blob/main/data/4_theorystrings_humans_extended_withcategories.csv

12. For the full list of person names see: https://github.com/theory-in-dh/conceptual_forays/blob/main/data/reference.theorists.full.csv

13.For a full list of all person last names see: https://github.com/theory-in-dh/conceptual_forays/blob/main/data/reference.theorists.partial.csv

Bibliography

Alvarado, R. C. (2019). “Digital Humanities and the Great Project: Why We Should Operationalize Everything and Study Those Who Are Doing so Now.” In *Debates in the Digital Humanities* 2019. Edited by Matthew K. Gold and Lauren F. Klein 5. Minneapolis: University of Minnesota Press.

Anderson, C.(2008). "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *Wired*, 2008. https://www.wired.com/2008/06/pb-theory/(accessed 27 February 2009).

Castle, G. (2008). *The Blackwell Guide to Literary Theory*. Blackwell guides to literature. Malden, Mass.: Blackwell.

Drucker, J. (2012). "Humanistic Theory and Digital Scholarship." In *Debates in the Digital Humanities*. Edited by Matthew K. Gold. Minneapolis, MN: University of Minnesota Press.

Gao, J., Nyhan, J., Duke-Williams, O., & Mahony, S.(2018). Visualising the digital humanities community: A comparison study between citation network and social network. *Digital Humanities 2018: Puentes-Bridges*. DH2018, Mexico City.

González, J. E., Jacobson, E., García, L. G., & Kujman, L. B.(2021). Measuring Canonicity: Graduate Reading Lists in Departments of Hispanic Studies. Journal of Cultural Analytics, 1(2).

Gutiérrez De la Torre, S. E.(2020). Bibliotecas y Humanidades Digitales en América Latina. *Revista de Humanidades Digitales*, *5*, pp. 113–131. https://doi.org/10.5944/rhd.vol.5.2020.27826(accessed 10 March 2022).

Hall, M. M.(2019). DH is the study of dead Dudes. In: Digital Humanities: multimedial & multimodal Konferenzabstracts (Sahle, Patrick ed.), Trier, Germany, pp. 111–113.

Heuser, R. & Le-Khac, L.(2012). "A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method." Stanford Literary Lab Pamphlets (4).

Kleymann, R., Niekler, A. & Burghardt, M.(currently under review). "Conceptual Forays: A Corpus-based Study of 'Theorytellings' in Digital Humanities Journals." In *Theorytellings: Epistemic Narratives in the Digital Humanities*. Special Issue of Cultural Analytics. Edited by Manuel Burghardt, Jonathan D. Geiger, Rabea Kleymann, Mareike Schumacher.

Luhmann, J. & Burghardt, M. (2021). "Digital Humanities – A Discipline in its Own Right? An Analysis of the Role and Position of DH in the Academic Landscape." In *Journal of the Association for Information Science and Technology*(JASIST), Special issue on Digital Humanities. https://asistdl.onlinelibrary.wiley.com/doi/pdfdirect/10.1002/asi.24533(accessed 12 April 2022).

Müller, M. C. (2017). Semantic author name disambiguation with word embeddings. In International conference on theory and practice of digital libraries. Springer, Cham, pp. 300-311.

Pawlicka-Deger, U.(2020). "The Laboratory Turn: Exploring Discourses, Landscapes, and Models of Humanities Labs." *Digital Humanities Quarterly*14, no. 3. http://www.digitalhumanities.org/dhq/vol/14/3/000466/000466.html(accessed 10 March 2022).

Poole, A. H.(2017). The conceptual ecology of digital humanities. In *Journal of Documentation*, 73, 91–122. https://doi.org/10.1108/ JD-05-2016-0065 (accessed 10 March 2022).

Rivkin, J. & Ryan M.(eds.) (2007). *Literary Theory: An Anthology*. 2. ed. Malden, Mass: Blackwell Publ.

Saklofske, J.(2016). "Digital Theoria, Poiesis, and Praxis: Activating Humanities Research and Communication Through Open Social Scholarship Platform Design." S*cholarly and Research Communication*7, no. 2, 1–16. DOI:10.22230/src.2016v7n2/3a252

Selden, R., Widdowson, P. & Brooker, P.(2006). *A Reader's Guide to Contemporary Literary Theory*. 5. ed., Harlow, England: Pearson Longman.

Out of the Slaughterhouse: The Birth of the Modern Detective Story Corpus

Hammond, Adam
adam.hammond@utoronto.ca
University of Toronto, Canada

Stern, Simon
simon.stern@utoronto.ca
University of Toronto, Canada

Introduction

We introduce the Birth of the Modern Detective Story (BMDS) Corpus, a new dataset for exploring the evolution of detective fiction during a crucial period of its development. It currently comprises 380 detective stories published between 1890 and 1920, provided in full text with relevant metadata and richly annotated for 61 categories related to the types of crimes, clues, and evidence represented. The project explores in more systematic fashion some of the work begun in Franco Moretti's "The Slaughterhouse of Literature" (2000). All data is freely accessible and open-access. We explain our motivations for developing the dataset, describe the dataset, explain our theoretical discoveries, and lay out some research trajectories.

Motivations

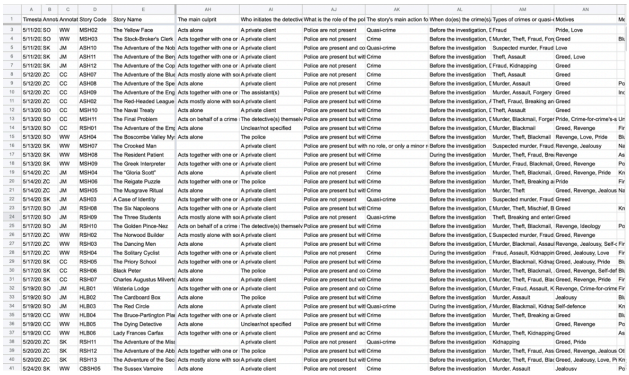
Moretti's "Slaughterhouse of Literature" asks a fascinating question: Why did the Sherlock Holmes stories "survive" when so many contemporaneous "competitors" were forgotten? It comes to an interesting conclusion: that their success *cannot* be explained by their use of what Moretti calls "decodable" clues — which, he finds, are neither exclusive to the Holmes stories nor consistently employed in them.

However promising, Moretti's approach is hampered by theoretical and methodological limitations as well as by its limited and opaque corpus. His categories of clues — necessary, visible, decodable — are poorly defined. The corpus of texts he investigates is small, inaccessible, and not described in detail. He performs two experiments: the first on a set of "about twenty" stories he calls "very narrow" and "haphazard" in selection; the second on 108 detective stories published in *The Strand* in the 1890s, the titles of which are not provided. Given its opaque corpus and imprecise terms, it is impossible to verify the article's claims.

Description of the BMDS Corpus

The BMDS corpus offers several methodological advances over "Slaughterhouse," and makes available all texts and annotations available. Whereas Moretti focuses haphazardly on the 1890s, we look at the years 1890–1920, during which the genre's conventions are agreed to have consolidated (Humphreys 2017). We aim to include *all* detective stories published in English during this period. At present, we have 380 stories. We began with those available for free in accessible data formats (now complete), moved next to those that can be purchased (underway), and will then ingest those that must be scanned and OCR'ed.

The BMDS Corpus is being assembled as follows. Starting in May 2021, we employed eleven separate student annotators. Working in pairs, students read a story and fill out a form asking them 61 questions, with terms set out in the Annotation Guidelines (Hammond et al. 2021). These include questions about the number, gender, and role of the story's detectives, assistants, victims, and culprits; the types and motivations of crimes investigated in the story; the types of clues and evidence present in the story; whether the crime is solved; and how subjectively satisfying the story is. This data is recorded in tabular form. The Corpus also includes Dublin Core metadata for all stories (380) and authors (20) in the corpus. Further, it includes every story in plain text form.



	A	B	C	D	E	AH	AI	AJ	AK	AL	AM	AN	
1	Timesta	Annot	Annot	Story Code	Story Name	The main culprit	Who initiates the detective	What is the role of the pol	The story's main action to	When do(es) the crime(s)	Types of crimes or quasi-c	Motives	Me
3	5/11/20	SO	WW	MSH02	The Yellow Face	Acts alone	A private client	Police are not present	Quasi-crime	Before the investigation, t	Fraud	Pride, Love	
4	5/11/20	SO	WW	MSH03	The Stock-Broker's Clerk	Acts together with one or	A private client	Police are not present	Crime	Before the investigation, t	Murder, Theft, Fraud, For	Greed	Bl
5	5/11/20	SK	JM	ASH10	The Adventure of the Not	Acts together with one or	A private client	Police are present and co	Quasi-crime	Before the investigation	Suspected murder, Fraud	Love	
6	5/11/20	SK	JM	ASH11	The Adventure of the Ber	Acts together with one or	A private client	Police are present but will	Crime	Before the investigation	Theft, Assault	Greed, Love	
7	5/11/20	SK	JM	ASH12	The Adventure of the Cop	Acts together with one or	A private client	Police are not present	Crime	Before the investigation, t	Fraud, Kidnapping	Greed	
8	5/12/20	ZC	CC	ASH07	The Adventure of the Blue	Acts mostly alone with so	A private client	Police are not present	Crime	Before the investigation	Theft, Assault	Greed	
9	5/12/20	ZC	CC	ASH08	The Adventure of the Spec	Acts alone	A private client	Police are not present	Crime	Before the investigation, t	Murder, Assault	Greed	Po
10	5/13/20	ZC	CC	ASH09	The Adventure of the Eng	Acts together with one or	The assistant(s)	Police are present but will	Crime	Before the investigation	Murder, Assault, Forgery	Greed	Inc
11	5/13/20	ZC	CC	ASH02	The Red-Headed League	Acts mostly alone with so	A private client	Police are present but will	Crime	Before the investigation, t	Theft, Fraud, Breaking an	Greed	
12	5/13/20	SO	CC	MSH10	The Naval Treaty	Acts alone	A private client	Police are present but will	Crime	Before the investigation, t	Theft, Assault	Greed	
13	5/13/20	SO	CC	MSH11	The Final Problem	Acts on behalf of a crime	The detective(s) themselv	Police are present but will	Crime	Before the investigation, t	Murder, Blackmail, Forger	Pride, Crime-for-crime's-s	Un
14	5/13/20	SO	WW	RSH01	The Adventure of the Em	Acts alone	Unclear/not specified	Police are present but will	Crime	Before the investigation, t	Murder, Blackmail	Greed, Revenge	Fir
15	5/13/20	SO	WW	ASH04	The Boscombe Valley My	Acts alone	The police	Police are present but will	Crime	Before the investigation	Murder, Theft, Blackmail	Revenge, Love, Pride	Bl
16	5/13/20	SK	WW	MSH07	The Crooked Man	Acts alone	A private client	Police are present but with no role, or only a minor r		Before the investigation	Suspected murder, Fraud	Revenge, Jealousy	Na
17	5/13/20	SK	WW	MSH08	The Resident Patient	Acts together with one or	A private client	Police are present but will	Crime	During the investigation	Murder, Theft, Fraud, Bre	Revenge	As
18	5/13/20	SK	WW	MSH09	The Greek Interpreter	Acts together with one or	A private client	Police are present but will	Crime	Before the investigation, t	Murder, Fraud, Blackmail,	Greed, Revenge	Po
19	5/14/20	ZC	JM	MSH04	The "Gloria Scott"	Acts alone	A private client	Police are not present	Crime	Before the investigation, t	Murder, Theft, Blackmail,	Greed, Revenge, Pride	Kn
20	5/14/20	ZC	JM	MSH06	The Reigate Puzzle	Acts together with one or	The police	Police are present but will	Crime	Before the investigation	Murder, Theft, Breaking a	Pride	Fir
21	5/14/20	ZC	JM	MSH05	The Musgrave Ritual	Acts alone	A private client	Police are present but will	Crime	Before the investigation	Murder, Theft	Greed, Revenge, Jealous	Na
22	5/14/20	SK	JM	ASH03	A Case of Identity	Acts together with one or	A private client	Police are not present	Quasi-crime	Before the investigation	Suspected murder, Fraud	Greed	
23	5/17/20	SO	JM	RSH08	The Six Napoleons	Acts together with one or	A private client	Police are present but will	Crime	Before the investigation, t	Murder, Theft, Mischief, B	Greed	Kn
24	5/17/20	SO	JM	RSH09	The Three Students	Acts mostly alone with so	A private client	Police are not present	Quasi-crime	Before the investigation	Theft, Breaking and enter	Greed	
25	5/17/20	SO	JM	RSH10	The Golden Pince-Nez	Acts on behalf of a crime	The detective(s) themselv	Police are present but will	Crime	Before the investigation	Murder, Theft, Blackmail,	Revenge, Ideology	Po
26	5/17/20	ZC	WW	RSH02	The Norwood Builder	Acts mostly alone with so	A private client	Police are present but will	Crime	Before the investigation, t	Suspected murder, Fraud	Greed, Revenge	
27	5/17/20	ZC	WW	RSH03	The Dancing Men	Acts alone	A private client	Police are present but will	Crime	Before the investigation, t	Murder, Blackmail, Assaul	Revenge, Jealousy, Self-r	Fir
28	5/17/20	ZC	WW	RSH04	The Solitary Cyclist	Acts together with one or	A private client	Police are not present	Crime	During the investigation	Fraud, Assault, Kidnappin	Greed, Jealousy, Love	Fir
29	5/17/20	SK	CC	RSH05	The Priory School	Acts together with one or	A private client	Police are present but will	Crime	Before the investigation, t	Murder, Blackmail, Kidnap	Greed, Jealousy, Pride	Bl
30	5/17/20	SK	CC	RSH06	Black Peter	Acts alone	The police	Police are present and co	Crime	Before the investigation	Murder, Theft, Blackmail,	Greed, Revenge, Self-def	Bl
31	5/17/20	SK	CC	RSH07	Charles Augustus Milverto	Acts alone	A private client	Police are present but will	Crime	Before the investigation, t	Murder, Theft, Fraud, Bla	Greed, Revenge, Pride	Fir
32	5/19/20	SO	JM	HLB01	Wisteria Lodge	Acts together with one or	A private client	Police are present and co	Crime	Before the investigation, t	Murder, Fraud, Assault, K	Revenge, Crime-for-crime	Fir
33	5/19/20	SO	JM	HLB02	The Cardboard Box	Acts alone	The police	Police are present but will	Crime	Before the investigation, t	Murder, Assault	Jealousy	Bl
34	5/19/20	SO	JM	HLB03	The Red Circle	Acts alone	A private client	Police are present but will	Quasi-crime	During the investigation, t	Murder, Blackmail, Kidnap	Self-defence	Kn
35	5/19/20	CC	WW	HLB04	The Bruce-Partington Pla	Acts together with one or	A private client	Police are present but will	Crime	Before the investigation	Murder, Theft, Breaking a	Greed	Bl
36	5/19/20	CC	WW	HLB05	The Dying Detective	Acts alone	Unclear/not specified	Police are present but will	Crime	Before the investigation	Murder	Greed, Revenge	Po
37	5/19/20	CC	WW	HLB06	Lady Frances Carfax	Acts together with one or	A private client	Police are present and ac	Crime	Before the investigation, t	Murder, Theft, Kidnapping	Greed	As
38	5/20/20	ZC	SK	RSH11	The Adventure of the Mis		A private client	Police are not present	Quasi-crime	Before the investigation	Kidnapping	Greed, Pride	
39	5/20/20	ZC	SK	RSH12	The Adventure of the Abb	Acts together with one or	The police	Police are present but will	Crime	Before the investigation	Murder, Theft, Fraud, Ass	Greed, Revenge, Jealous	Ol
40	5/20/20	ZC	SK	RSH13	The Adventure of the Sec	Acts mostly alone with so	A private client	Police are present but will	Crime	Before the investigation, t	Murder, Theft, Fraud, Bla	Greed, Jealousy, Love, Pr	Kn
41	5/24/20	SK	WW	CBSH05	The Sussex Vampire	Acts alone	A private client	Police are not present	Crime	Before the investigation	Murder, Assault	Jealousy	Po

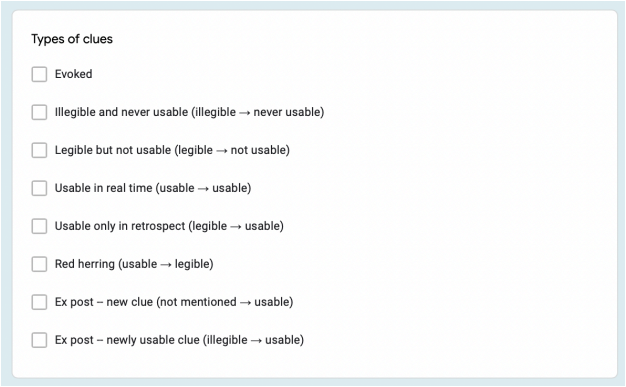
Screen shot from tabular data recorded for individual stories.

Theoretical Contributions

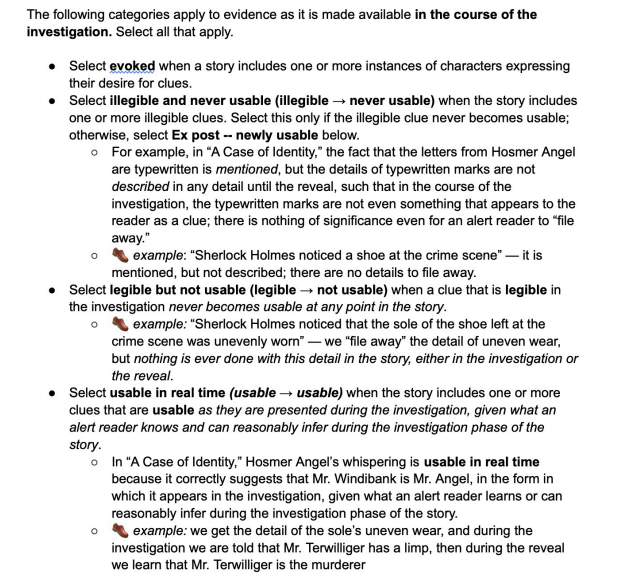
The BMDS Corpus makes several important theoretical advances over "Slaughterhouse"; most notably, it provides several new categories of clues, and it introduces the distinction between the "investigation" and "reveal" phases of detective stories.

Our Guidelines clearly distinguish between four types of clue: evoked, illegible, legible, and usable. For instance, a legible clue is defined as one that is "presented in sufficient detail to appear to the reader as a clue" whereas a usable clue is "a legible clue that leads an alert reader in the direction of the correct solution to the crime" (this distinction is not present is Moretti).

Our understanding of clues depends not merely on whether clues are *present* but how they are *used* in the story's plot. We divide detective fiction plots into two parts: the "investigation phase," during which the crime is actively investigated; and the "reveal phase," during which the detective presents their solution. Our clue categories depend on how particular clues *appear to the reader* in each of these two phases; for example, we distinguish between a clue that is "legible but not usable" (one whose legibility does not point an alert reader in the direction of the correct solution) versus one that is "usable in real time" (one that does). (See Figs 2 and 3).



Screen shot from the input form annotators use to enter findings. This is one of 61 questions, which asks the annotator to input the types of clues employed in the given story.



The following categories apply to evidence as it is made available **in the course of the investigation.** Select all that apply.

- Select **evoked** when a story includes one or more instances of characters expressing their desire for clues.
- Select **illegible and never usable (illegible → never usable)** when the story includes one or more illegible clues. Select this only if the illegible clue never becomes usable; otherwise, select **Ex post -- newly usable** below.
  - For example, in "A Case of Identity," the fact that the letters from Hosmer Angel are typewritten is *mentioned*, but the details of typewritten marks are not *described* in any detail until the reveal, such that in the course of the investigation, the typewritten marks are not even something that appears to the reader as a clue; there is nothing of significance even for an alert reader to "file away."
  - *example*: "Sherlock Holmes noticed a shoe at the crime scene" — it is mentioned, but not described; there are no details to file away.
- Select **legible but not usable (legible → not usable)** when a clue that is **legible** in the investigation *never becomes usable at any point in the story.*
  - *example:* "Sherlock Holmes noticed that the sole of the shoe left at the crime scene was unevenly worn" — we "file away" the detail of uneven wear, but *nothing is ever done with this detail in the story, either in the investigation or the reveal.*
- Select **usable in real time *(usable → usable)*** when the story includes one or more clues that are **usable** *as they are presented during the investigation, given what an alert reader knows and can reasonably infer during the investigation phase of the story.*
  - In "A Case of Identity," Hosmer Angel's whispering is **usable in real time** because it correctly suggests that Mr. Windibank is Mr. Angel, in the form in which it appears in the investigation, given what an alert reader learns or can reasonably infer during the investigation phase of the story.
  - *example:* we get the detail of the sole's uneven wear, and during the investigation we are told that Mr. Terwilliger has a limp, then during the reveal we learn that Mr. Terwilliger is the murderer

Screen shot from Annotation Guidelines corresponding to the choice in Fig. 2.

Research Applications

Our fuller specification of the types and functions of clues makes it possible to refine and revise Moretti's claims about the counterintuitive result he discerned, concerning the haphazard and inconsistent use of "decodable" clues. Changes over time that reflect our distinction between "legible" and "usable" clues suggest that mystery writers *did* in fact come to rely more heavily on clues during this period — but not necessarily clues planted for readers to use. Detectives were increasingly required to solve crimes by reasoning on the basis of trace evidence, and this pattern shows a marked change during the period surveyed here.

Another possible application involves the gender dynamics that animate these stories. Although both authors and detectives were usually male, a considerable number of women wrote in this genre, sometimes with female detectives. Exploring the gender of the authors, detectives, victims, suspects, and culprits may allow for insights into styles of detection, types of clues, varying emphasis on different types of crimes, and the like.

The "investigation" / "reveal" distinction may facilitate various kinds of discoveries relating to some of the issues above and other discoveries involving the stories' structure and effects. Does the ratio of the investigation phase to the reveal phase vary with types of clues, or types of crimes? Over time, do we see that ratio stabilize, as writers come to perceive preferences among readers? Does the language of the reveal phase exhibit distinctive features that appear to enhance reader satisfaction?

Bibliography

Hammond, A., Stern, S., Colclough, Z., Côté, C., Maharaj, A., Maleshev, M, Michielin, J., Oh, S., Kim, S., Selvaraj, M., Wen, W. (2021). BMDS Annotation Guidelines. https://tinyurl.com/bmdsguidelines (accessed April 28, 2020).

Humphreys, A. (2017). British Detective Fiction in the 19th and Early 20th Centuries. *Oxford Research Encyclopedia: Literature*.

Moretti, F. (2000). The Slaughterhouse of Literature. *Modern Literature Quarterly* 61(1), pp. 207–227. 2000.

Voices Speaking To and About One Another: Introducing the Project Dialogism Novel Corpus

Hammond, Adam
adam.hammond@utoronto.ca
University of Toronto, Canada

Vishnubhotla, Krishnapriya
vkpriya@cs.toronto.edu
University of Toronto, Canada

Mohammad, Saif M.
saif.mohammad@nrc-cnrc.gc.ca
National Research Council Canada, Ottawa, ON, Canada

Hirst, Graeme
gh@cs.toronto.edu
University of Toronto, Canada

Introduction

We introduce a new dataset for the computational analysis of novels: the Project Dialogism Novel Corpus (PDNC). The PDNC currently consists of 22 novels in which all quotations are identified and annotated for speaker, addressee(s), and characters mentioned. PDNC is by an order of magnitude the largest corpus of its kind. Each novel is annotated manually by a pair of annotators using customized software we developed. In addition to releasing the dataset itself alongside this paper, we are also releasing the custom annotation software we developed (including the source code) along with our annotation guidelines. In the discussion section, we present two applications of the PDNC from our own research: quote attribution and emotion dynamics. We argue that the PDNC will promote a more nuanced and accurate view of novelistic discourse; whereas much research currently envisions the novel as expressing the voice of the *author*, the PDNC presents novels as a polyphonic fabric of *characters'* voices.

Overview of the Project Dialogism Novel Corpus

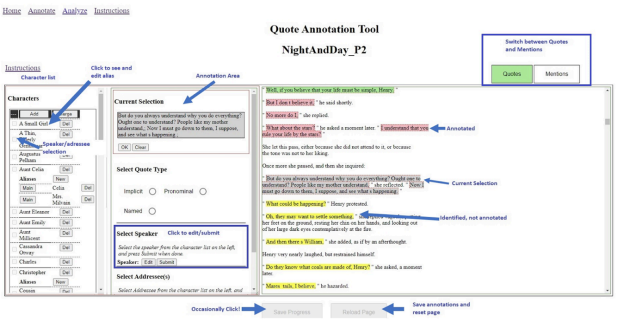
The PDNC currently consists of 22 novels (see Table 1). In selecting novels, our aim has been to annotate texts in a variety of genres (literary fiction, children's literature, detective fiction, and science fiction are represented); from the LitBank (REF #1) and QuoteLi (REF #15) corpora, to facilitate comparison and validation; of broad interest to a variety of scholars while still relevant to our group's interest in stylistic diversity and dialogism. Further, we have chosen to annotate multiple novels by Jane Austen, in order to facilitate comparative analysis of a single author's oeuvre (Austen was chosen because she is included in all existing corpora).

The annotation workflow proceeds as follows. First, the novel is pre-processed in GutenTag (Brooke et al. 2015); from this, a provisional character list is built and likely quotations are identified. Next, the novel is manually annotated in our customized software (see Figure 1). This is done separately by two annotators. Working from our guidelines (Hammond et al. 2021), annotators select each quotation, then identify the speaker, addressee, and anyone mentioned in the quotation (whether by name or pronoun). Annotators also identify the referring expression for each quotation, as well as the quotation type: explicit (quotations in which the referring expressions give the character's name; for example, "said Emma"), pronominal (pronoun given; "she said"), or implicit (no referring expression). Once both annotators have completed their work, their annotations are compared for any discrepancies. The annotators then meet to resolve any disagreements, in what we call a "consensus exercise." Once comparison shows no disagreement between annotations, the novel is considered annotated.

The PDNC is by an order of magnitude the largest corpus of its kind (see Table 2). The largest previous corpus of novels annotated in this manner is the QuoteLi corpus, which contains only three novels (*Pride and Prejudice* and *Emma*, both in PDNC; and Chekhov's *The Steppe*, not in PDNC). The LitBank corpus includes annotations for 100 novels, but only for a very small fraction of each is annotated (on average, only 2,000 words). The Columbia Quoted Speech Attribution Corpus consists of six texts, two of which are compilations of short stories, but they are only partly annotated for quote attribution.

Novel	Author	# Tokens	# Quotations	# Speakers	# Addressees	# Mentions
Emma (1815)	Jane Austen	188131	2116	16	4169	412
Northanger Abbey (1817)	Jane Austen	90208	1017	16	1601	336
Persuasion (1817)	Jane Austen	96667	702	24	1678	133
Pride and Prejudice (1813)	Jane Austen	143804	1708	27	4200	2121
Sense and Sensibility (1811)	Jane Austen	139968	1545	20	3172	164
Alice's Adventures in Wonderland (1865)	Lewis Carroll	34339	1048	32	3544	84
The Man Who Was Thursday (1908)	G. K. Chesterton	69246	1357	28	4129	568
The Awakening (1899)	Kate Chopin	58925	738	18	991	981
The Mysterious Affair at Styles (1921)	Agatha Christie	72602	2226	28	7366	379
The Sign of the Four (1890)	Sir Arthur Conan Doyle	51790	891	18	1697	296
The Sport of the Gods (1902)	Paul Laurence Dunbar	50013	830	34	1370	814
The Gambler (1866; 1910)	Fyodor Dostoevsky (Trans. C. J. Hogarth)	73557	1068	20	2495	832
Howards End (1910)	E. M. Forster	136812	3131	46	5060	836
A Room with a View (1908)	E. M. Forster	83383	1989	27	3588	836
The Sun Also Rises (1926)	Ernest Hemingway	89123	3316	41	5904	2315
Daisy Miller (1879)	Henry James	26607	725	10	1209	456
Anne of Green Gables (1908)	Lucy Maud Montgomery	123465	1779	25	2412	165
A Handful of Dust (1934)	Evelyn Waugh	89070	2617	70	3628	381
The Age of Innocence (1920)	Edith Wharton	120052	1600	31	2430	714
The Invisible Man (1897)	H. G. Wells	60033	1274	29	1759	209
The Picture of Dorian Gray (1891)	Oscar Wilde	95631	1501	31	2356	1228
Night and Day (1919)	Virginia Woolf	199450	2800	38	4137	217

Table 1.
PDNC: Tokens, quotations, speakers, total # of addressees recorded, total # of mentions



Screen shot from our custom annotation software.

Corpus	*Columbia Quoted Speech Attribution Corpus (2010)*	*He et al. (2013)*	*QuoteLi (2017)*	*LitBank (2020)*	*PDNC (2021)*
# Texts	6	3	3	100	22
# Annotated Quotations	3176	1901	3103	1765	35978

Table 2.
Comparison of PDNC with previous quotation attribution corpora

Research Applications

The research applications of the PDNC are multiple, extending well beyond the boundaries of our own research interests. Yet our own research serves to demonstrate some of its possible uses.

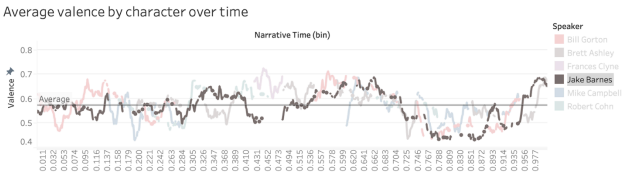
We began developing the PDNC primarily to test our quote attribution system (Hammond et al. 2020). The corpus has proven essential to this work, allowing us to compare our systems against state-of-the-art systems like QuoteLi and the BERT-based system in the latest release of BookNLP (see Table 3).

		State-of-the-art			Ours	
Novel	Author	# Quotations	Muzny et al.	BookNLP	# Quotations	Accuracy
Emma (1815)	Jane Austen	1981	0.584	0.602	2102	**0.676**
Northanger Abbey (1817)	Jane Austen	1008	0.415	0.449	998	**0.702**
Persuasion (1817)	Jane Austen	631	0.569	0.605	663	**0.685**
Pride and Prejudice (1813)	Jane Austen	1696	0.563	0.561	1683	**0.682**
Sense and Sensibility (1811)	Jane Austen	1409	0.579	0.58	1529	**0.642**
Alice's Adventures in Wonderland (1865)	Lewis Carroll	951	0.894	0.92	1004	**0.959**
The Man Who Was Thursday (1908)	G. K. Chesterton	1300	0.506	0.523	1295	**0.798**
The Awakening (1899)	Kate Chopin	595	0.481	0.476	708	**0.689**
The Mysterious Affair at Styles (1921)	Agatha Christie	2121	0.162	0.209	2187	**0.60**
The Sign of the Four (1890)	Sir Arthur Conan Doyle	702	0.463	0.464	855	**0.677**
The Sport of the Gods (1902)	Paul Laurence Dunbar	766	0.403	0.373	761	**0.570**
The Gambler (1866; 1910)	Fyodor Dostoevsky	933	0.23	0.251	1038	**0.754**
Howards End (1910)	E. M. Forster	3087	0.524	0.504	3015	**0.626**
A Room with a View (1908)	E. M. Forster	1936	0.505	0.521	1954	**0.614**
The Sun Also Rises (1926)	Ernest Hemingway	2183	**0.777**	0.76	3163	0.738
Daisy Miller (1879)	Henry James	720	0.688	0.686	713	**0.833**
Anne of Green Gables (1908)	Lucy Maud Montgomery	1723	0.714	0.748	1722	**0.880**
A Handful of Dust (1934)	Evelyn Waugh	2231	0.531	**0.537**	2467	0.522
The Age of Innocence (1920)	Edith Wharton	1481	0.189	0.196	1535	**0.683**
The Invisible Man (1897)	H. G. Wells	1190	0.629	0.646	1207	**0.812**
The Picture of Dorian Gray (1891)	Oscar Wilde	1384	**0.676**	0.621	1445	0.669
Night and Day (1919)	Virginia Woolf	2783	0.58	0.631	2728	**0.689**

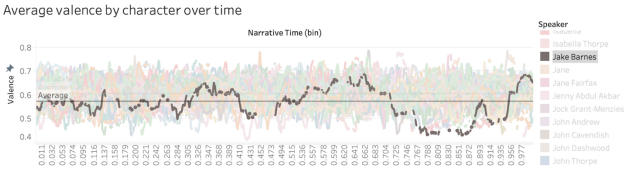
Table 3.
A comparison of performance of our latest quote attribution system vs. QuoteLi vs. BookNLP. Numbers reported are accuracy scores; best scores are bolded.

Perhaps the largest aim of PDNC is to reorient computational work away from conceiving novels as undifferentiated lumps of text attributed solely to their authors — but rather as complex fabrics of differentiated voices speaking to and about one another, mediated by a narrator. In the paper introducing the tool GutenTag (Hammond and Brooke 2017), one of our authors used a rudimentary version of PDNC to rebut Matthew Jockers's (2013) claim that female novelists generally write about stereotypically feminine themes. By looking at character voices *within* novels, however, rather than attributing all the novel's text to its author, we demonstrated that it was female *characters* who discussed these themes — and that Jockers's results were a secondary consequence of the fact that female authors tended to include far more female characters in their works. By allowing researchers to look *within* novels and analyze novels through the voices that make them up, PDNC will shift research away from mistaken assumptions and conclusions like Jockers's.

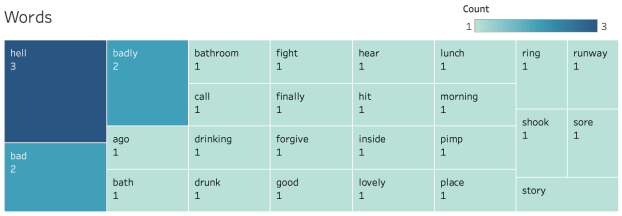
Our work on "emotion dynamics" — the study of change in emotional states over time — presents another example of new research enabled by the PDNC. Sentiment analysis is among the richest and most vital areas of computational literary research today. Yet major work seeking to plot novels' sentiment trajectories remains limited by the necessity of assuming a single source for all words: the author (Elsner 2012, Mohammad 2011, Jockers 2014, Reagan 2016). In a pioneering essay on "emotion dynamics" in films, Hipson and Mohammad (2021) show the benefits of considering *individual characters'* emotional trajectories. This approach enables researchers to determine each character's "home base" (typical emotional range) as well as their emotional variability and the speed at which they regulate variations. We are currently working to apply this approach to the novels in PDNC (Figures 2–4 show the emotional trajectory of Jake Barnes in Ernest Hemingway's *The Sun Also Rises*, revealing that this reputedly taciturn character in fact experiences one of the most extreme emotional troughs (in terms of valence) of any character in PDNC). We are using this approach to test whether characters' emotion dynamics track with familiar literary-critical categories such as flat vs. round characters (Forster 1927). We are also investigating the extent to which emotional trajectories are gendered, and whether male or female authors are more likely to create characters that diverge from gender norms.



Emotion dynamics trajectory, valence only, for characters in Ernest Hemingway's The Sun Also Rises. Jake Barnes's emotional trajectory is highlighted; the trough three-quarters of the way through the novel (~76%-87%) occurs during and after his fight with Robert Cohn at the Fiesta.



Emotion dynamics, valence only, for all characters in PDNC. Jakes Barnes's trajectory (highlighted) is extreme in the context of the characters in our corpus.



Emotion words (with frequency count) used by Jake Barnes during trough (76%-87% portion of novel)

Bibliography

Bamman, D., Popat, S., and Shen, S. (2019). An annotated dataset of literary entities. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2138-2144.

Brooke, J., Hammond, A., and Hirst, G. (2015). GutenTag: an NLP-driven tool for digital humanities research in the Project Gutenberg corpus. *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pp. 42-47.

Elsner, M. (2012). Character-based kernels for novelistic plot structure. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 634-644.

Elson, D. K. and McKeown, K. R. (2010). Automatic attribution of quoted speech in literary narrative. *Twenty-Fourth AAAI Conference on Artificial Intelligence*. 2010.

Forster, E. M. (1927). *Aspects of the Novel.* New York: Harcourt, Brace, and Company.

Hammond, A. and Brooke, J. (2017). GutenTag: A User-Friendly, Open-Access, Open-Source System for Reproducible Large-Scale Computational Literary Analysis. *Proceedings of the Digital Humanities 2017 Conference*, pp. 246–249.

Hammond, A., Vishnubhotla, K., and Hirst, G. (2020). The Words Themselves: A Content-Based Approach to Quote Attribution. *Proceedings of the Digital Humanities 2020 Conference.*

Hammond, A., Vishnubhotla, K., Duarte, L., Oh, S., Pajovic, J., and Siegal, B. (2022). Annotation Guidelines for the Project Dialogism Novel Corpus. https://tinyurl.com/quoteattribution (accessed April 28, 2022).

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 174-184.

Brooke, J, and Hirst, G. (2013). A multi-dimensional Bayesian approach to lexical style. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 673-679.

He, H., Barbosa, S., and Kondrak, G. (2013). Identification of speakers in novels. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1312-1320.

Hipson, W.E., and Mohammad, S. E. (2021). Emotion dynamics in movie dialogues. *PloS one* vol. 16(9).

Jockers, M. (2013). *Macroanalysis: Digital Methods and Literary History* (University of Illinois Press).

Jockers, M. (2014). "A novel method for detecting plot." http://www.matthewjockers.net/2014/06/05/a-novel-method-for-detecting-plot/ (accessed April 28, 2022).

Mohammad, S. (2011). From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales. *Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (LaTeCH).

Muzny, F., Fang, M., Chang, A., and Jurafsky, D. (2017). "A two-stage sieve approach for quote attribution. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 460-470.

Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., and Sheridan Dodds, P. (2016). The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science* 5(31), pp. 1–12.

Realizing a multilingual tool for best legal and ethical practices in DH research: The ELDAH Consent Form Wizard as a model for community-driven internationalization

Hannesschläger, Vanessa
vanessa.hannesschlaeger@gmail.com
Austrian National Library, Austria

Kuzman-Slogar, Koraljka
koraljka@ief.hr
Institute of Ethnology and Folklore Research, Zagreb, Croatia

Scholger, Walter
walter.scholger@uni-graz.at
University of Graz, Austria

In May 2018, the European Union General Data Protection Regulation (GDPR) came into effect. This rigorous legislation on data protection and privacy, among many other things, significantly changed the way research with data from living persons is conducted since it introduced a number of obligations and restrictions to ensure a high level of protection of individuals' rights. The research community has been struggling with fulfilling these new obligations and desperately requires training, knowledge and tools to support their research under the new circumstances.

In reaction to this demand, the working group Ethics and Legality in Digital Arts and Humanities (ELDAH) of the European research infrastructure consortium DARIAH-EU developed the ELDAH Consent Form Wizard (CFW) , a tool that provides researchers, cultural heritage institutions and research infrastructures with GDPR-compliant consent form templates for DH research purposes. With the help of this tool, valid consent to processing personal data according to the current European legal regulations can be gathered. The CFW offers support and guidance with three main scenarios:

1. gathering data from and/or about living people for research purposes
2. communicating through mailing lists or other (digital) communication media
3. collecting data and/or consent from participants as the host of an academic event

The tool, which was developed based on community-driven input by (mostly) European DH researchers and administrators, has been well received and is broadly used even beyond the scope of its legal application, since colleagues from South-Africa, New Zealand and India have also been using this tool as a means to transparently provide information to and gather informed consent from their research subjects: not in order to fulfill a legal obligation but to establish a high and transparent ethical standard in their own research practices despite legal differences.

Other than fulfilling legal and ethical community demands, the development of the CFW tool also addresses another important topic in the dissemination of DH methods and contents: the goal to create a multilingual resource. A prominent and successful example for efforts at multilingual tool development and internationalisation are the TEI Publisher and the activities of the TEI working group on Internationalisation.

In a legal and ethical context, multilingualism is particularly important not only in terms of increased accessibility but also as a means to create trust and accountability between researchers and their data subjects when gathering and processing personal data for research purposes. Participants in interviews, surveys and the likes need to fully understand their rights and the purposes for which their data are collected in order to make an informed decision about their participation. Therefore, having such resources available in their native language should be considered best practice. The - originally English - ELDAH CFW has already addressed this need with translations into Croatian, Italian and German. Additional translations into French, Greek, Slovene and even Hindi are already in progress, while translations into Spanish and Japanese are being planned.

For the translation of the UI, a dedicated translation interface was developed. With the help of this interface, CFW translators can internationalize the tool to their national languages line by line, field by field, button by button. Since the tool provides not only building blocks for a consent form, but also aims to educate its users with larger informative text passages as they complete the online form, a fully automated, dictionary-based translation approach was considered too inaccurate and therefore inadequate, especially since translations of legal contexts and formulations have to be accurate and reliable. While the translation of the webform used to collect the information from the individual user of the CFW worked very well, the consequently automatically generated template for the consent form ended up with some room for improvement. Furthermore, a purely participant-driven

approach to translations in a field which is subject to regular revisions and changes (because of legislation changes) poses significant problems regarding sustainability and accuracy.

The proposed paper will discuss two important aspects, presenting the CFW as a showcase and starting point: a) the importance of tools that provide support to the global research community regarding aspects of data protection (or legal and ethical research standards in general), and b) the challenges, ways and benefits of developing such a tool as a multilingual resource.

We would like to highlight the potential of our internationalization approach as a template for other projects, e.g. for the development of a multilingual library of terms that are commonly used on comparable interfaces. In addition, we would like to critically discuss issues that can arise in the course of translation - especially in the case of legal tech where portability and compatibility with national legislations are essential. Keeping a multilingual tool sustainable and accurate presents a tremendous challenge: we will explore how this could be achieved as a collaborative, community-driven effort and invite colleagues to share their own expertise in implementing multilingual resources.

Bibliography

Asef, E. / Wagner, C. / Lee, M. / Nowak, S. (2019): *Workshop Report Non-Latin Scripts in Multilingual Environments: research data and digital humanities in area studies*. https://blogs.fu-berlin.de/bibliotheken/2019/01/18/workshop-nls2018/

Bermúdez Sabel, H. / Cayless, H. / Meneses, L. / Nagasaki, K. / Rio Riande, G. / Scholger, M. (2019): *Communicating the TEI Across Linguistic and Cultural Boundaries*. https://trianglesci.org/2019/07/17/communicating-the-tei-to-a-multilingual-user-community

DARIAH-EU Working Group Ethics and Legality in Digital Arts and Humanities (2020): *The ELDAH Consent Form Wizard*. https://consent.dariah.eu/

e-editiones (2017): *TEI publisher*. https://teipublisher.com/

European Union (2019): *Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC*. https://eur-lex.europa.eu/eli/dir/2019/790/oj

Kamocki, P. (2021): *Handouts on the processing of personal data for the purposes of language research and archiving of language resources under the General Data Protection Regulation*. https://doi.org/10.14618/ids-pub-10695

Kamocki, P. / Ketzan, E. / Wildgans, J. (2018): *Language Resources and Research Under the General Data Protection Regulation*. http://nbn-resolving.de/urn:nbn:de:bsz:mh39-97562

CCVG Data: A Unique, Curated, and Searchable Chinese Village Dataset for Chinese Study Scholars

He, Daqing
dah44@pitt.edu
University of Pittsburgh, United States of America

Ma, Rongqian
rom77@pitt.edu
University of Pittsburgh, United States of America

Zheng, Ruoyun
ruz38@pitt.edu
University of Pittsburgh, United States of America

Zhang, Haihui
haihuiz@pitt.edu
University of Pittsburgh, United States of America

Chinese local gazetteers are unique primary resources with a long history in China. According to Endymion Wilkinson, "[Gazetteers] form one of the most important sources for the study of Chinese history in the past one thousand years" (2000, p. 154). During the past 30 years, North American-centered Chinese research has developed rapidly. Scholars recognize that there is still a need for more ways to pursue systematic research, thus requiring more reliable information sources. Although some data on Chinese rural areas are available, they are scattered across different resources at varying levels of accessibility.

The Contemporary Chinese Village Gazetteer Data Project (CCVG Data) is a project initiated and conducted by the East Asian Library (EAL) of the University of Pittsburgh Library System (ULS). CCVG Data, the first of its kind in the world, is designed to establish a dataset of significant humanities and social science value based on the East Asian Library's extensive collection of village gazetteers. Village gazetteers report quantitative and qualitative data for China's most basic administrative unit, covering a series of topics such as local history, genealogy, economics, education, politics and management, public health, etc. Data at this level of detail can only be found

in village gazetteers. In 2018, CCVG Data was awarded a one-year, $35,100 Pitt Seed Project grant supported by the University of Pittsburgh's Office of the Chancellor. In August 2019, CCVG Data project completed its first stage as a pilot project and made the first 500 villages' data public for teaching and research of Chinese studies. By the end of September 2021, about 1,500 villages' data have been extracted, entered, and opened to the public. To enhance the accessibility of data as well as to improve user experience, in January 2020, the CCVG team initiated a collaborative project with the School of Computing and Information at the University of Pittsburgh. Two important tasks are under way to increase online accessibility to CCVG data: First, storing the extracted village gazetteer data into a database, and second, providing powerful search interfaces for scholars.

In the presentation, we will briefly introduce the background and rationales of CCVG Data project, such as its data extraction, quality control, and data dictionary, and then we will focus on our ongoing efforts to design and construct the database and search interface. Currently, CCVG data is stored in a MySQL database, where 38 tables were identified for storing the data related to the villages' 12 thematic topics such as gazetteer information, village basis, natural environment, natural disaster, ethnic groups, economy, and education. Two online search interfaces are provided to CCVG data: single village search for basic search and multiple village search for advanced search. The single village search enables scholars to access and download information related to one village on all 12 topics. The multiple village search provides advanced search capabilities, where several villages can be selected with filters on province, city, and county. The 12 topics are presented as a tree structure for the scholars to browse and select, which forms more filters to specify the relevant CCVG data. Scholars can make single-year selections or year range inputs to impose further filters. The resulting data is downloadable for the scholars. Our initial assessments on the search interfaces indicated their usefulness to the scholars.

The CCVG Data project has received much attention from scholars and researchers in a variety of disciplines. Inquiries on using CCVG Data have been received from many disciplines of humanities and social studies such as religion, education, economic, family planning and public health, local management, etc. Following the overview of the CCVG Data interface and its usage among scholars, in this presentation, we will also discuss the value of CCVG Data project and the collaboration model which has contributed to the progress of the project.

From books to a dataset, CCVG Data is an experimental project with significant value to the humanities and social sciences. It supports Chinese studies in fields such as politics, economics, sociology, environmental science, history, and public health, and proves to be a meaningful exploration in Digital Humanities (DH). More specifically, with the rapid development of East Asian DH, the CCVG Data project marks a milestone in this emerging field from multiple perspectives (Vierthaler, 2020). On the one hand, CCVG Data project demonstrates libraries' leadership roles in DH work and represents a deeper, more well-rounded collaboration model between librarians, humanities scholars, social scientists, and computer scientists. The collaboration covers various aspects and nearly every step of the DH work, such as data collection and processing, database and platform design, user research, and programming and development of digital tools. As collaboration becomes a well-recognized topic that matters significantly for the success of DH research, the CCVG Data project suggests diverse roles that information professionals can serve in DH scholarship (Poremski, 2017; Richardson & Eichmann-Kalwara, 2017; Risam et al., 2017). On the other hand, CCVG Data project contributes to DH scholarship with infrastructural innovation. Unlike the extensive bibliographic databases, tools, and textual corpora available for East Asian studies, the CCVG Data project focuses on numeric datasets and leverages computational methods (e.g., searchable interface, visual analytic tools) to facilitate users' and researchers' interaction with the data.

Bibliography

CCVG Data website: www.chinesevillagedata.library.pitt.edu

Poremski, M. D. (2017). Evaluating the landscape of digital humanities librarianship. College & Undergraduate Libraries, 24(2–4), 140–154. https://doi.org/10.1080/10691316.2017.1325721

Richardson, H. A. H., & Eichmann-Kalwara, N. (2017). Process and collaboration: Assessing digital humanities work through an embedded lens. College & Undergraduate Libraries, 24(2–4), 595–615. https://doi.org/10.1080/10691316.2017.1336145

Risam, R., Snow, J., & Edwards, S. (2017). Building an ethical digital humanities community: Librarian, faculty, and student collaboration. College & Undergraduate Libraries, 24(2–4), 337–349. https://doi.org/10.1080/10691316.2017.1337530

Vierthaler, P. (2020). Digital humanities and East Asian studies in 2020. History Compass, 18(11), e12628.Wilkinson, E. P. (2000). Chinese History: A Manual, Revised and Enlarged. Cambridge, Mass: Harvard University Asia Center for the Harvard-Yenching Institute.

Wilkinson, E. P. (2000). Chinese History: A Manual, Revised and Enlarged. Cambridge, Mass: Harvard University Asia Center for the Harvard-Yenching Institute.

Structuring the Management of Research Data - Reflections on Requirements and Service Concepts in Research Data Management in the Humanities

Helling, Patrick
patrick.helling@uni-koeln.de
Data Center for the Humanities (DCH), University of Cologne, Germany

Introduction and Background

Research data is a key element of ongoing scientific progress (Bryant et al. 2017). Making research data findable, accessible, interoperable, and reusable in the sense of the FAIR Principles (Wilkinson et al. 2016) is a substantial aspect of good research practice (DFG 2019). Fulfilling this goal is a central task for both researchers and research data management (RDM) competence centers.

For defining service structures of these RDM competence centers, different guidelines and information bases are used (cf. HRK 2014 and 2015; Akers and Doty 2013; Mathiak et al. 2019; Vock 2019). [1] In addition, formal models have been developed to describe RDM service structures (cf. Rans and Whyte 2017; Hiemenz and Kuberek 2019; Quin et al. 2017; Lemaire et al. 2020; Herterich et al. 2019).

Most of these guidelines, models and information are based on either top-down recommendations, quantitative surveys which are highly fault-prone or description models that focus on the perspective of service points and infrastructures.

However, RDM should be steered by the daily necessities of the researchers (e.g., table 1). A formal model that describes the structures, processes, and conditions at the core of RDM practice from both the perspectives of the researcher and the data manager does not seem to exist yet. Nonetheless, our knowledge of RDM needs to be structured to guarantee the quality of research.

In this paper, I will present a new approach to the definition of RDM requirements of researchers and service workflows within the Humanities by qualitatively analyzing RDM counseling protocols. These protocols were produced during RDM consultation sessions with researchers by an RDM competence center in Germany for documentation reasons.

Identification and analysis of RDM-requirements and workflows

90 semi-structured and anonymized RDM counseling protocols were used for my study. With a qualitative analysis and an inductive definition of categories (Mayring 2015), I identified 48 categories of RDM requirements (see Table 1). The RDM counseling protocols were clustered according to these categories.

Total number of identified RDM requirements			
Proposal	27	Visualization	3
Archiving	25	Accessibility	3
Website	17	Internal access	2
Database	16	Licenses	2
Storage	13	Tools	2
Counseling	9	Acquisition	2
Data management plans	9	Survey	2
Generic research data management	9	Mediation	2
Cooperation	9	Consolidation	2
Software	8	Analysis	1
Metadata	7	Accompanying support	1
Data publication	7	RDM-support during the project duration	1
Backup	5	Operation of software	1
Digitization	5	Data handling	1
Policies	5	Longtail data	1
Legal issues	5	Concept of sustainability	1
Research data infrastructure	4	Reusability	1
Storage during the project duration	4	OCR	1
Publication of other outputs	3	Project management	1
Data collection	3	Support of the project	1
Data presentation	3	Training	1
Standards	3	Snapshots	1
Technical support	3	Letter of intent	1
Availability	3	Virtual research environments	1

Table 1:
Total number of identified RDM requirements.

Four main goals were pursued in the study of these clusters:

1. Identification and formalization of relevant information describing the categories.
2. Development of formal description models of the requirements.
3. Identification and formalization of documented workflows.
4. Mapping of the workflows to define comprehensive recommendations.

Two RDM requirements will serve as examples to illustrate my work.

Requirement category “Proposal support”

In 27 documented counselling sessions researchers asked for support while writing a project proposal. This category was differentiated into two sub-requirements: (a)

review of a chapter on RDM in a proposal ("Review section of text") and (b) writing of a chapter on RDM for a proposal ("Write section of text") by the center.

In the first case many different RDM aspects played a role in the reviewing process (see Figure 1), while in the second case fewer aspects were discussed during the consultation (see Figure 2). This might be due to the fact that writing a section of text on RDM by a competence center is a more targeted and structured process, while sections of text written by researchers not expert in RDM may be more uncertain and fault prone.



Figure 1:
Topics and aspects while reviewing a section of text.



Figure 2:
Topics and aspects while writing a section of text.

Nevertheless, in both cases the archiving of data was a main RDM aspect (see Figures 3 and 4).

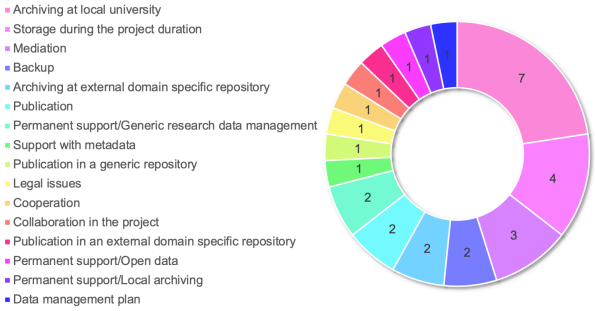


Figure 3:
Frequencies of topics and aspects while reviewing a section of text.

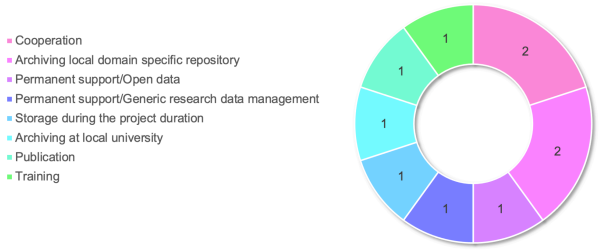


Figure 4:
Frequencies of topics and aspects while writing a section of text.

Prototypical RDM counselling workflow - "Proposal support"

Also, in the case of the cluster "Proposal support" I differentiated between prototypical workflows for the sub-requirements "Review section of text" and "Write section of text" (Figures 5 and 6).

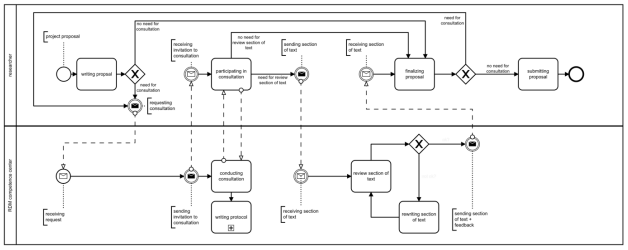


Figure 5:
Prototypical workflow "Review section of text".

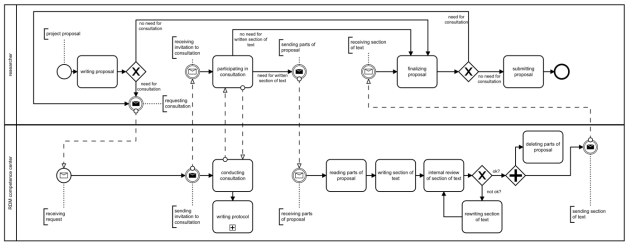


Figure 6:
Prototypical workflow "Write section of text".

For the review of a section of text, the competence center corrects and annotates the text sent by the researcher. In the case of the sub-requirement "Write section of text", the researcher shares relevant parts of the proposal to allow the members of the center to draft a section of text. Clearly, the information on the project is immediately deleted after the conclusion of the workflow for data protection reasons.

The workflows usually present more complexities, as I will show in my presentation.

Requirement category “Archiving”

In the 25 cases belonging to the requirement of archiving research data, various information could be identified both describing the RDM requirement and influencing how it should be dealt with (see Figure 7).

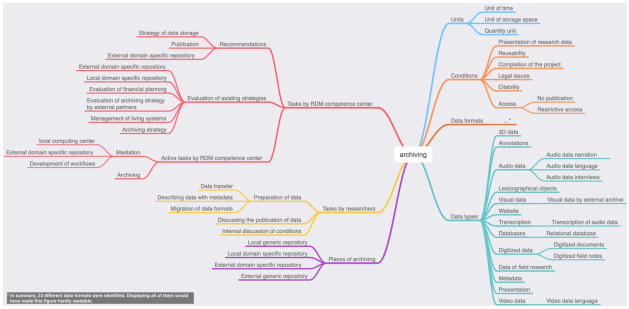


Figure 7:
Topics and aspects while supporting an archiving process.

Besides concrete information on the objects to be archived, the identification of conditions for the archiving process by the researchers revealed to be key in the process of archiving, showing RDM awareness among the researchers.

In addition, various types of recommendations by the RDM competence center as well as concrete tasks for both the center and the researchers could be identified.

Prototypical RDM counselling workflow - “Archiving”

To support the researchers in the archiving process, the RDM competence center generally looks for a suitable repository to store the research data (see Figure 8). If there is no fitting domain- or data specific repository the center seeks a more generic solution.

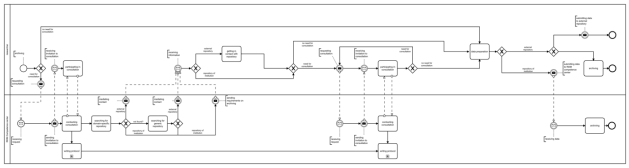


Figure 8:
Prototypical workflow “Archiving”.

If there is a suitable solution at the center, conditions for archiving are communicated to the researchers, who prepare their research data and submit it to the center for archiving. If an external solution is recommended, the RDM competence center transmits the necessary information to the researchers, who get in touch with the external repository, prepare and submit the data (or proceed with self-archiving).

Conclusion

With the described approach, one gets closer to a description of the RDM requirement landscape based on the daily needs of researchers. The analysis allows for qualitative and quantitative descriptions of RDM requirements based on real counseling sessions. It gives a better understanding of the nature of specific RDM requirements, their conditions and influencing factors. Workflows to deal with specific RDM requirements are modeled prototypically RDM competence centers. Specific competences can be identified and integrated accordingly.

In my talk, I will present the descriptions of RDM requirements and corresponding workflows comprehensively. Specifically, I will show dependencies and mappings between different conditions and information within the RDM requirement clusters to concretely describe the RDM landscape. Finally, I will focus on the mapping between the RDM requirement models and their prototypical workflows.

Bibliography

Akers, K. G., and Doty, J. (2013). Disciplinary differences in faculty research data management practices and perspectives. *The International Journal of Digital Curation*, Volume 8, Issue 2. DOI: 10.2218/ijdc.v8i2.263.

Bryant, R., Lavoie, B., and Malpas, C. (2017). A Tour of the Research Data Management (RDM) Service Space. The Realities of Research Data Management, Part 1. 2017. Dublin, Ohio: OCLC Research. DOI: .

DFG, Deutsche Forschungsgemeinschaft (2019). Guidelines for Safeguarding Good Research Practice. Code of Conduct. DOI: 10.5281/zenodo.3923601.

Herterich, P., Davidson, J., Whyte, A., Molloy, L., Matthews, B., and Kayumbi Kabeya, G. (2019). D6.1 Overview of needs for competence centres. *FAIRsFAIR*, 2019. DOI: 10.5281/zenodo.3549790.

Hiemenz, B., and Kuberek, M. (2019). Strategischer Leitfaden zur Etablierung einer institutionellen Forschungsdaten-Policy. DOI: 10.14279/depositonce-8412.

HRK, Hochschulrektorenkonferenz (2015). Empfehlung der 19. Mitgliederversammlung der HRK am 10. November 2015 in Kiel. Wie Hochschulleitungen die Entwicklung des Forschungsdatenmanagements steuern können. Orientierungspfade, Handlungsoptionen, Szenarien. Online: https://www.hrk.de/uploads/tx_szconvention/Empfehlung_Forschungsdatenmanagement__final_Stand_11.11.2015.pdf [last request 15th of November 2021].

HRK, Hochschulrektorenkonferenz (2016). Empfehlung der 16. Mitgliederversammlung der HRK am 13. Mai 2014 in Frankfurt am Main. Management von Forschungsdaten - eine zentrale strategische Herausforderung für Hochschulleitungen. Online: https://www.hrk.de/uploads/tx_szconvention/HRK_Empfehlung_Forschungsdaten_13052014_01.pdf [last request 15th of November 2021].

Lemaire, M., Gerhards, L., Kellendonk, S., Blask, K. and Förster, A. (2020). Das DIAMANT-Modell 2.0. Modellierung des FDM-Referenzprozesses und Empfehlungen für die Implementierung einer institutionellen FDM-Servicelandschaft (eSciences Working Papers, 05). Trier. DOI: 10.25353/ubtr-xxxx-f5d2-fffb.

Mayring, P. (2015). Qualitative Inhaltsanalyse, Grundlagen und Techniken. 12. Auflage, Weinheim und Basel: Beltz Verlag.

Mathiak, B., Metzmacher, K., Helling, P. and Blumtritt, J. (2019). The Role Of Data Archives In The Humanities At The University Of Cologne. *DH 2019 Conference*, 8-12 July 2019, Utrecht University. DOI: 10.34894/geqeko.

Qin, J., Crowston, K. and Kirkland, A. (2017). Pursuing Best Performance in Research Data Management by Using the Capability Maturity Model and Rubrics. *Journal of eScience Librarianship;6(2): e1113*. DOI: 10.7191/jeslib.2017.1113.

Rans, J. and Whyte, A. (2017). Using RISE, the Research Infrastructure Self-Evaluation Framework, v.1.1. Edinburgh: Digital Curation Centre. Online: www.dcc.ac.uk/resources/how-guides [last request 17th of November 2021].

Vock, R. with the participation of Gerlach, R., Hesse, B., Colomb, J., Steiner, P., Schröter, A., Hiltscher, A., Prinz, T. and König-Ries, B. (2019). Evaluation der FDM-Beratung 2019 – Evaluation des Beratungsangebots der Kontaktstelle Forschungsdatenmanagement (KS FDM) an der Friedrich-Schiller-Universität Jena. *Bericht 4.3. eeFDM-Projekt (BMBF)*, Jena. Online: https://www.db-thueringen.de/receive/dbt_mods_00040382 [last request 15th of November 2021].

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Bonino da Silva Santos, L., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. In: *Scientific Data 3*, Article number: 160018. DOI: 10.1038/sdata.2016.18.

Notes

1. see for example: Listing of surveys on Research Data Management, Online: https://www.forschungsdaten.org/index.php/Umfragen_zum_Umgang_mit_Forschungsdaten_an_wissenschaftlichen_Institutionen (last request: 15th of November 2021).

Making Research Data FAIR. Seriously? - Reflections on Research Data Management in the Computational Literary Studies

Helling, Patrick
patrick.helling@uni-koeln.de
University Würzburg, Germany

Jung, Kerstin
kerstin.jung@ims.uni-stuttgart.de
University Würzburg, Germany

Pielström, Steffen
pielstroem@biozentrum.uni-wuerzburg.de
University Würzburg, Germany

Introduction

Computational Literary Studies (CLS) are an evolving, interdisciplinary field of research combining research

questions from the traditional field of Literary Studies with methods and technologies from Computer Sciences and Computational Linguistics. The German Research Foundation (DFG) is funding a priority program to foster the ongoing evolution of the field and the development and establishment of innovative computational methods in literary studies: [1] The priority program comprises eleven research projects in Germany and Switzerland and one central project (Pielström et al. 2021) for improving the interdisciplinary exchange between the projects and developing a common and domain-specific research data management (RDM) strategy.

Research data produced within the CLS is, similarly to many other disciplines in the humanities, heterogeneous (Pempe 2012). The management of this research data is a key element of scientific progress (Bryant, Lavoie & Malpas 2017) and a substantial aspect of good research practice (DFG 2019); in this respect, a major landmark are the FAIR Principles (Wilkinson et al. 2016). Within the central project of the program, we interviewed all projects (Helling et al. 2020) with regard to their discipline specific methods and approaches as well as the data and software they both use and produce during their research. We analysed the interviews qualitatively and quantitatively. The results of the survey (Helling et al. 2021) are used to develop and establish a common RDM strategy for the whole priority program to meet the FAIR Principles and enhance the sustainable findability, accessibility, interoperability and reusability of the data and outcomes of the projects.

In this paper we present our experience in RDM within the program. We will illustrate both pragmatic RDM solutions and major barriers in making research data FAIR. We will show that these barriers are intrinsic in the discipline itself.

Pragmatic Solutions and Barriers in Making Research Data FAIR in the CLS

We recommended Zenodo, which meets the FAIR principles, as a fallback solution for storing the outputs of the projects within the program, because overarching domain-specific infrastructures within the CLS are very rare. In fact, most of our projects were already using Zenodo for publishing outputs that do not fit into other infrastructures. However, the research data of the program is also stored and published in institutional, generic and domain-specific repositories (see Figure 1 and 2).

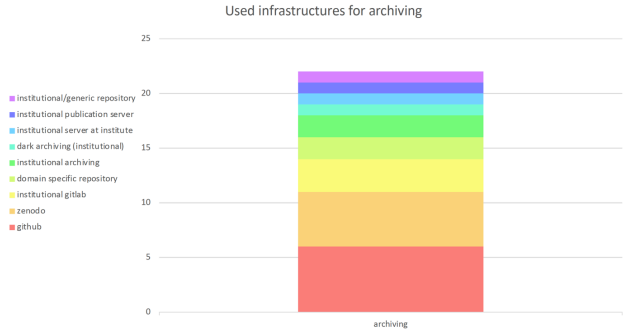


Figure 1:
Used infrastructures for archiving within the program.

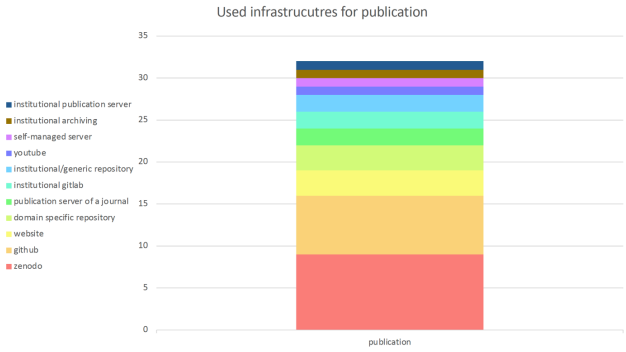


Figure 2:
Used infrastructures for publication within the program.

Technical Perspectives on FAIR

From a technical perspective, repositories used within the program address the FAIR Principles fairly well. Regarding the findability and accessibility of research data, outputs of the program are usually registered or indexed in searchable resources (F4), which are accessible via standardized, open, and universally implementable protocols (A1/A1.2). Besides Zenodo, some of the used infrastructures support the assignment of a persistent identifier (PID) (F1). Moreover, supplied metadata is often based on the generic DataCite scheme (F2). [2] In addition, most of the repositories and infrastructures offer the definition of generic licenses and the possibility of making the research data gradationally accessible (R1.1).

Domain-Specific Perspectives on FAIR

A large set of primary data in the priority program is beyond copyright by age, thus licenses for reuse in research and education are unproblematic. The remaining smaller set of data can be restricted by personal rights (studies) or individual copyright negotiations with authors, publishers, or libraries. In the community, there is a large interest to

make data as accessible as possible and provide secure licenses (R1.1). Still for some aspects there are no clear solutions or test cases, such as the context necessary in derived formats (Schöch et al. 2020). The matter is regularly discussed in a working group on copyright within the program.

Schemes to capture provenance metadata (R1.2) are still evolving in the DH domains (cf. Gärtner et al. 2018). In contrast to e.g., the life sciences the objective is not a fully automatically reproducible workflow, but an equal treatment of automatic and manual steps which pose the domain-specific challenge. Individual documentation is available as well as commit histories from GitHub repositories. [3] So overall these aspects are evolving along the lines of FAIR.

Regarding interoperability (I1/2/3) and the relevance of attributes (R1) and standards (R1.3) the CLS requires to distinguish between resource-related metadata, content-related metadata, and data from annotations. Resource-related metadata is and can be based on DataCite (cf. F2 above), whereas more domain-specific metadata ranges from information on time period, genre and author uncertainty over technical settings like encoding (e.g., TEI variants) to the overall application of very different methodologies (see Figure 3) which comprises the existence of specific annotation layers as well as different segmentation schemes. Thereby neither content-related metadata nor annotation categories can come with a fixed, agreed on, common vocabulary since these categories are an integral part of the research (outcome) itself. [4] Still this vocabulary poses the basis for search, exploration and the FAIRness of the built resources.

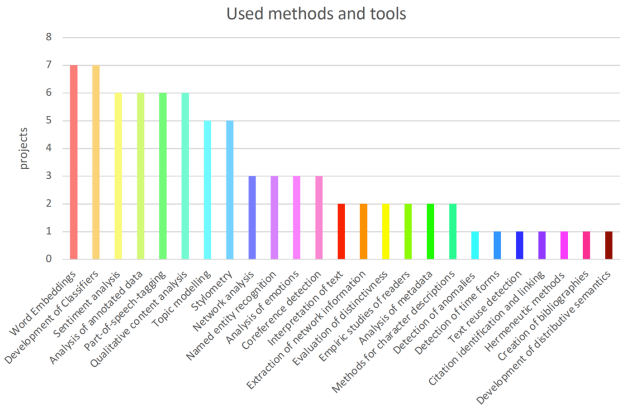


Figure 3:
Used methods and tools within the program.

Conclusion

While different domain-specific as well as generic/ institutional repositories meet the FAIR principles at least partially, Zenodo (also in combination with GitHub) is the closest infrastructure to the FAIR principles which is used in the context of the program. Nevertheless, it is still difficult to make research data stored in generic/ institutional repositories findable for specific research communities, especially since a domain-specific metadata scheme is missing. Moreover, a common vocabulary for such a scheme possibly cannot exist without losing relevant content, differing between research fields within the domain. This problem is of course addressed by some more domain-specific infrastructures but still a comprehensive and domain-specific description model for the CLS is not existing.

In sum, without domain-specific metadata schemes, sustainable infrastructures and guiding legal implementations of copyright handling for the CLS, the FAIR principles can hardly be addressed in their entirety in this research domain.

Currently, pragmatic RDM seems to be the only way to meet the FAIR principles at least partially and to do effective RDM for the research community. In our talk, we will present more of our pragmatic RDM solutions and illustrate our approach to improve FAIRness of CLS research data for the CLS community. In this regard, a pragmatic approach for harvesting the heterogenous achievements of the program will be discussed. Finally, we will address specific RDM requirements for the CLS for fulfilling the FAIR principles and plead for a more domain-specific and measurable interpretation and implementation of the FAIR principles.

Bibliography

Bryant, R., Lavoie, B. and Malpas, C. (2017). A Tour of the Research Data Management (RDM) Service Space. The Realities of Research Data Management, Part 1. Dublin, Ohio: OCLC Research. DOI: 10.25333/C3PG8J.

DFG - Deutsche Forschungsgemeinschaft (2019). Guidelines for Safeguarding Good Research Practice. Code of Conduct. DOI: 10.5281/zenodo.3923602.

Eckart, K. and Heid, U. (2014). Resource interoperability revisited. *Proceedings of the 12th edition of the KONVENS conference Hildesheim*, Germany, pp. 116-26. URN: https://nbn-resolving.org/urn:nbn:de:gbv:hil2-opus-2725.

Gärtner, M., Hahn U., and Hermann, S. (2018). Supporting Sustainable Process Documentation. Rehm G., Declerck T. (eds) *Language Technologies for the Challenges of the Digital Age*. GSCL 2017. Lecture Notes in Computer Science, vol 10713. Springer, Cham DOI: 10.1007/978-3-319-73706-5_24.

Gius, E., Meister, J. C., Meister, M., Petris, M., Bruck, C., Jacke, J., Schumacher, M., Gerstorfer, D., Flüh, M., and Horstmann, J. (2018-2021). CATMA. Concept DOI: 10.5281/zenodo.1470118.

Helling, P., Jung, K., Reiter, N. and Pielström, S. (2020). Interviewleitfaden zur FDM-Bestandsaufnahme im Schwerpunktprogramm Computational Literary Studies. DOI: 10.5281/zenodo.4269639.

Helling, P., Jung, K. and Pielström, S. (2021). Disziplinspezifisches Forschungsdatenmanagement - FDM-Bedarfserfassung in den Computational Literary Studies. *FORGE 2021 Konferenz: Forschungsdaten in den Geisteswissenschaften - Mapping the Landscape - Geisteswissenschaftliches Forschungsdatenmanagement zwischen lokalen und globalen, generischen und spezifischen Lösungen (FORGE 2021)*, Cologne. DOI: 10.5281/zenodo.5379629.

Pempe, W. (2012). Geisteswissenschaften. In: Neuroth, H., Strathmann, S., Oßwald, A., Scheffel, R., Klump, J. and Ludwig, J. (eds), *Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme*. Boizenburg: Verlag Werner Hülsbusch, pp. 137-60.

Pielström, S., Helling, P. and Jung, K. (2021). Zentralprojekt des DFG-Schwerpunktprogramms Computational Literary Studies. *Program General Meeting*, virtuell. DOI: 10.5281/zenodo.5041338.

Schöch, C., Döhl, F., Rettinger, A., Gius, E., Trilcke, P., Leinen, P., Jannidis, F., Hinzmann M. and Röpke, J. (2020). Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. *Zeitschrift für digitale Geisteswissenschaften*. Wolfenbüttel. DOI: 10.17175/2020_006.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Bonino da Silva Santos, L., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. In: *Scientific Data 3*, Article number: 160018. DOI: 10.1038/sdata.2016.18.

Notes

1. https://dfg-spp-cls.github.io/ [last request: 24th of November 2021].
2. DataCite Metadata Schema 4.4, https://schema.datacite.org/meta/kernel-4.4/ [last request: 07th of December 2021].
3. GitHub, https://github.com [last request: 07th of December 2021].
4. Regarding annotations, Eckart and Heid (2014) argue for a separation of content-related interoperability and representation format-related interoperability. For the latter we found the projects in the priority program to agree on CATMA (Gius et al. 2018-2021) using its own TEI Export Format.

Building an OCR Pipeline for a Republican Chinese Entertainment Newspaper

Henke, Konstantin
konstantin.henke@pm.me
Heidelberg Centre for Transcultural Studies, Heidelberg University

Arnold, Matthias
arnold@hcts.uni-heidelberg.de
Heidelberg Centre for Transcultural Studies, Heidelberg University

In recent years, the digitisation of newspapers has made a lot of progress, and large national and international initiatives like Trove[1], Chronicling America[2], Europeana Newspapers[3], Impresso[4], NewsEye[5], Oceanic Exchanges[6], OCR-D[7], Deutsches Zeitungsportal[8], and Living with Machines[9] emerged that are building up on and going beyond sheer digitisation, venturing into various areas of content analysis (Oberbichler et al, 2021). Also, the outcomes of these initiatives are usually provided online with open access, and publications increasingly follow the FAIR principles (Wilkinson et al, 2016). However, most of the textual content covered is printed in Latin script languages, and to a large degree the analytical systems rely on linguistic features like word boundaries, digital lexica, or tagged corpora.

Responding to this from an Asian perspective, i.e. looking at materials from regions where non-Latin scripts prevail, the situation is different. In our case we are working

with newspapers from Republican China. Although there are some projects working on historical Chinese newspapers (Stewart et al, 2020), results have so far rarely been published. Other initiatives provide their final results as commercial products. In general, a certain reluctance can be observed when it comes to publishing research methodologies, not to mention the open access sharing of ground truth, test corpora, or trained models (Arnold et al, forthcoming).

In our project we collected periodicals from the Republican era as image scans (Sung et al, 2014) and started OCR experiments to transform them into machine readable full text.

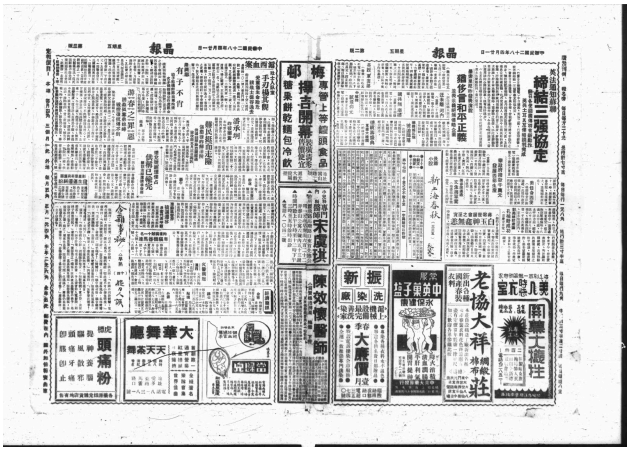


Fig. 1:
One of the 9385 fold scans of Jing bao

Looking closer at *Jing bao* 晶報 *(The Crystal)* (cf. Fig. 1), an entertainment newspaper that ran from 1919 to 1940, we soon learned that the key issue of OCR'ing the material actually lies in the page-level segmentation. We therefore started creating ground truth (GT) for geometrical data featuring semantically grouped bounding boxes with labels (article, image, advertisement, marginalia). We then used the resulting dataset to train dhSegment and have the network detect content areas on the folds (Arnold, forthcoming) (Fig. 2).

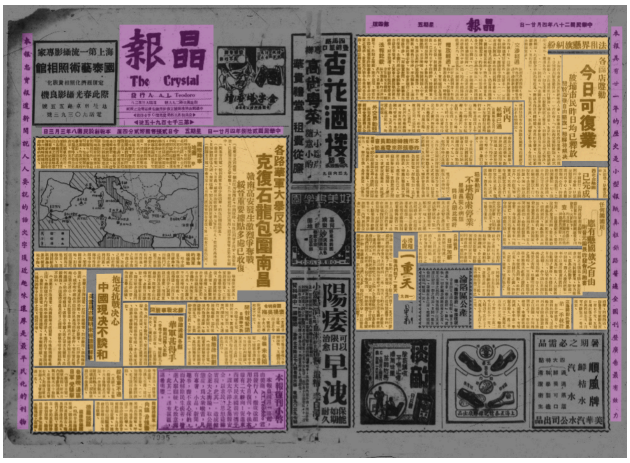


Fig. 2:
Automatic page segmentation results. Blocks with text content are shown in yellow.

Additionally, we created a text GT that not only covers all text in a machine readable local XML format, but also contains information about reading sequence running direction of the text. Based on this GT we were able to process a first set of manual crops, introducing a character segmentation method for grid-based printing layouts which produces over 90,000 labeled images of single characters (Henke, 2021). In this work, a GoogLeNet is trained as an OCR classifier on said character images after extensive pre-training on synthetical character image data created from font files. Additional error correction using language models yields an accuracy of 97.44%.

In our presentation we introduce our work on developing a document image processing pipeline currently focusing on Republican Chinese newspapers with complex layouts like the *Jing bao*. We will present the following concrete contributions:

1. A page-level segmentation approach (as seen in Fig. 2) yielding single text blocks.
2. An OCR pipeline taking single text blocks as input.

While Arnold (forthcoming) presented first promising experiments regarding (1), in this presentation we will concentrate on (2). Our evaluation metric for OCR output is the character error rate (CER) with regard to the ground truth annotation of every text block crop, which, based on the Levenshtein distance, is computed by:

$$CER = \frac{S + D + I}{L}$$

(S, D, I = number of substitutions, deletions, insertions; L = length of the reference sequence, i.e. corresponding GT text).

The character segmentation approach presented in Henke (2021) can however only process text blocks where characters are printed in a grid-like layout, which accounts for a very small portion of the *Jing bao*. Hence, there is a particular need for efficient character detection in less stable layout situations within text blocks, before passing single character images on to the actual OCR engine. As a baseline, we leverage the publicly available state-of-the-art OCR tool Tesseract (Smith, 2007) which provides out-of-the-box segmentation+recognition models even for vertically printed traditional Chinese. Tesseract however seems to struggle with the low input image resolution (~ 25x25 px per character) and overall inconsistent scan quality, leading to a very high CER of **47.85%** on the test set from Henke (2021).

To solve this issue, we use the readily-trained HRCenterNet from Tang et al. (2020) for character detection, and crop the bounding boxes to feed them into the GoogLeNet trained in Henke (2021). However, while our crops have a great variety of aspect ratios, the HRCenterNet expects at least nearly-squared rectangles. Hence, we cut the original images into 250x250 px tiles with a 50 px overlap (both horizontally and vertically, Fig. 3c). Bounding boxes (Fig. 3d) found in the overlapping sections are filtered during the non-maximum suppression (NMS) operation already included in the HRCenterNet pipeline (Fig. 3e).



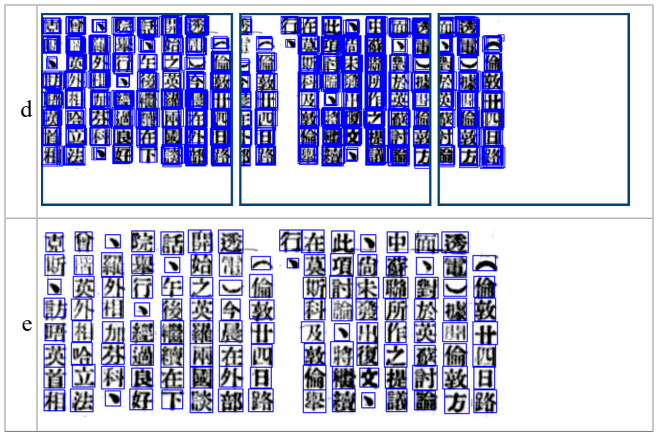


Fig 3:
a) original image, b) image after contrast enhancing, c) tiling with overlap, d) bounding boxes found by HRCenterNet before NMS, e) final result after reconnection of tiles and NMS

In addition, Fig. 4 shows how the HRCenterNet largely profits from contrast-enhancement (Fig 3b) during image pre-processing, especially for low-contrast input images.

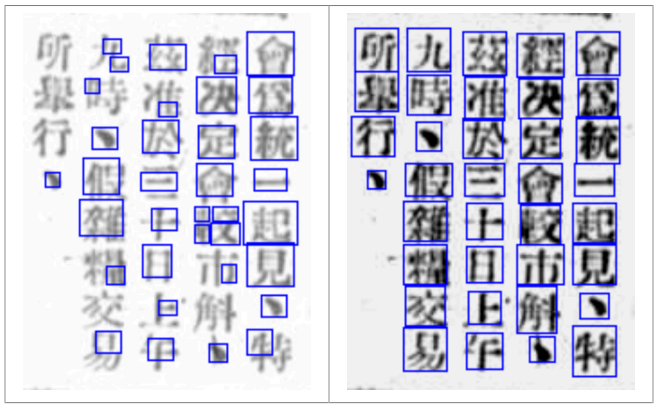


Fig 4:
Effects of contrast enhancement on character detection using HRCenterNet

Using the above method, the CER on the test set of Henke (2021) is reduced to **5.64%**.

In the presentation we will show how the results can be confirmed on a non-grid-based section of the corpus, for which we currently create GT annotations. We are confident that the additional pre-processing of crops and individual character images will help to further reduce the CER, and in combination with (1), yield a powerful document-level OCR pipeline for the *Jing bao* and similar Republican newspapers. This will not only open the door to further processing with the tools of Digital Humanities, but also

further contribute to FAIR-based work in the diverse Asian sphere.

Bibliography

Arnold, M. (2021). Multilingual Research Projects: Challenges for Making Use of Standards, Authority Files, and Character Recognition. Digital Studies/Le Champ Numérique, forthcoming. DOI: 10.11588/heidok.00030918 (preprint).

Arnold, M., Paterson, D. and Xie, J. (forthcoming). Procedural Challenges: Machine Learning Tasks for OCR of Historical CJK Newspapers. International Journal of Digital Humanities, Special issue on Digital Humanities and East Asian Studies. (manuscript accepted by special issue editors, currently under review by journal).

Henke, K. (2021). Building and Improving an OCR Classifier for Republican Chinese Newspaper Text. B.A. thesis, Heidelberg University. DOI: 10.11588/heidok.00030845

Oberbichler, S., Boros, E., Doucet, A., Marjanen, J., Pfanzelter, E., Rautiainen, J., Toivonen, H. and Tolonen, M. (2021). Integrated Interdisciplinary Workflows for Research on Historical Newspapers: Perspectives from Humanities Scholars, Computer Scientists, and Librarians. Journal of the Association for Information Science and Technology. DOI: 10.1002/asi.24565

Smith, Ray (2007). An Overview of the Tesseract OCR Engine. Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), pp. 629-633. DOI: 10.1109/ICDAR.2007.4376991

Stewart, S., 朱吟清 Zhu, Y., 吴佩臻 Wu, P., 赵薇 Zhao, W., Gladstone, C., Long, H., Detwyler, A. and So, R. J. (2020). 比较文学研究与数字基础设施建设：以“民国时期期刊语料库(1918-1949),基于PhiloLogic4”为例的探索 (Comparative Literature Research and Digital Infrastructure: Taking the ‘Republican China Periodical Corpus (1918-1949), Based on PhiloLogic 4’ as an Example). 数字人文 Digital Humanities, no. 1: 175–82. online version

Sung, D., Sun, L. and Arnold, M. (2014). The Birth of a Database of Historical Periodicals: Chinese Women’s Magazines in the Late Qing and Early Republican Period. TSWL 33, no. 2: 227–37. URL: https://www.jstor.org/stable/43653333

Tang, C., Liu, C. and Chiu, P. (2020). HRCenterNet: An Anchorless Approach to Chinese Character Segmentation in Historical Documents. IEEE International Conference on Big Data (Big Data). DOI: 10.1109/BigData50022.2020.9378051

Wilkinson, M. D., Dumontier, M., Aalbersberg, IJ. J., Appleton, G., Axton, M., Baak, A., Blomberg, N. et al. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. no. 1. Scientific Data 3, no. 1: 160018. DOI: 10.1038/sdata.2016.18

Notes

1. https://trove.nla.gov.au/newspaper/
2. https://chroniclingamerica.loc.gov/newspapers/
3. http://www.europeana-newspapers.eu/
4. https://impresso-project.ch/
5. https://www.newseye.eu/
6. https://oceanicexchanges.org/
7. https://ocr-d.de/en/
8. https://www.deutsche-digitale-bibliothek.de/newspaper
9. https://livingwithmachines.ac.uk/

Townsend & Sons, Account Book Manufacturer’s Business Guide and Works Manual: 19th Century Primary Manuscript Source and TEI Encoding

Hermsen, Lisa
lismariehermsen@gmail.com
Rochester Institute of Technology

Walker, Rebekah
rgwtwc@rit.edu
Rochester Institute of Technology

The session will describe the application of TEI guidelines to encode an idiosyncratic primary source manuscript–a volume from the collection of The William Townsend & Sons, Printers, Stationers, and Account Book Manufacturers, Sheffield UK (1830-1910). Volume 3, “Business Guide and Works Manual,” is a remarkable manuscript both for book history and cultural observations about unionization, gender roles, and credit/debit accounting. The extraordinary complexity of the manuscript’s structure requires that it be marked up to render a readable document. This project is using Text Encoding Initiative (TEI) Guidelines to create a digital edition, illuminating such key topics as the Townsend firm’s social networks, information on the men, women and apprentices who worked for the firm between 1830 and 1910, and the web of economic partners with whom the firm did business. The digital edition will be accompanied by documentation of editorial decisions about encoding that

will serve as a potential model for other digitized volumes in the Townsend collection, as well as other complex primary sources.

Volume 3 is the most rewarding volume, given its record of day-to-day activity in a stationery binding firm. Stationery binding is a type of underappreciated and lesser known book trade (Pickwoad, 2009). While the history of bookbinding most often refers to externally decorated bindings, stationery binding "satisfies" the inner workings of the useful book (Monk, 1912). William Townsend manufactured account books deliberately for double entry accounting. The account book itself, scholars note, changed practices from single entry to double entry accounting, a critical part of an emerging capitalist economy. Scholars agree that debit/credit bookkeeping coincided with the rise of factory industries, wage labor, and capital investment (Edwards, 1989) and that accounting in the nineteenth century was at least coincidently connected to the growth of modern capitalism (Geeson-White, 2012).

The Townsend manuscript is extraordinarily complex. It contains no obvious textual demarcations, no reliable pagination or progressive date notations. The manuscript may be considered a nineteenth-century common-place book, without the logical indices of the long tradition (Havens 2001). This index is alphabetical but subject headings are hardly informative, with entries like "thoughts and ideas" and "jottings." It is filled with notes tucked into margins and ephemera inserted in seemingly random places. There are marginal references to God, William Shakespeare, and Charles Dickens, as well as allusions to little- or unknown authors. Volume 3 contains hundreds of addresses for individual people and proprietor-named businesses. There are specific instructions for book binding, including lists of required materials and a recipe for glue. The manuscript is messy and the markup up requires flexible language, careful interpretation, and customized editing.

The Townsend volume records the cost of doing business, and so also belongs to a genre of documents—Historical Financial Records. It provides an opportunity to examine the relationship between financial artifacts and the worlds in which they were produced and embedded (Tomasek and Bauman, 2014). However, the financial accounts in this collection are recorded in ambiguous tabular form with in-text page references to nearly indecipherable price keys. For example, Townsend provides a "key" to determine the size of an account book. The formula is figured using imperial standards for the size of a sheet of paper (i.e. Foolscap) and quarto or octavo folds of the sheet and the number of sheets. This formula, along with the type of ruling and binding, provides the necessary numbers for the arithmetic that will determine the price of an account book.

We have encoded over 100 pages of this 400-page manuscript. The encoding has had to be innovative in the use of divisions and abstract blocks. The many tables have been difficult to represent, but the encoding reveals some logic. We have started to encode for labor in hours and wages, and in a process that allows for comparisons throughout the manuscript. We are identifying materials, with and without measurable units, and developing an adequate encoding practice. We will want to capture the methods of binding, if possible. We have yet to decipher the "keys" for calculating sizes and prices of paper, or even to understand the numbers on the page. The paper will draw upon work already underway to develop a robust markup for numbers and measurements (Kokaze, et al., 2020).

We are using several TEI projects as models for the Digital Townsend. Our introduction to TEI came through the Women Writers Project, which continues to frame our understanding of the way TEI guidelines can be used to markup complex and multi-layered documents. The Almanacks of Mary Moody Emerson offers insight on producing digital editions of unconventional (and long) volumes. Additional models, including the Folger Library's Early Modern Manuscripts Online, The Civil War Governors of Kentucky project, George Washington's Financial Papers project. The research and experimentation in the Accounts of the City of Basel is a model to consider the financial records in our volume and to determine whether the TEI guidelines are sufficient to meaningfully encode accounting practices, and the structuring of income and expenditure.

The Townsend collection was acquired by the Cary Graphic Arts Collection at the Rochester Institute of Technology as a part of The Bernard C. Middleton Collection of Books on the History and Practice of Bookbinding, one of the world's most comprehensive collections on the subject, including both historical and contemporary materials from (including but not limited to) England, France, India, Japan, Portugal, Sudan, and Turkey. All volumes are digitized and publicly available through RIT Libraries' Digital Collections.

Bibliography

Edwards, J.R. (1989). A History of Financial Accounting. United Kingdom: Routledge University Press.

Geeson-White, J. (2012). Double Entry: How the Merchants of Venice Created Modern Finance. W. W. Norton & Company.

Havens, E. (2001). Commonplace Books: A History of Manuscripts and Printed Books from

Antiquity to the Twentieth Century. New Haven: Yale University.

Kokaze, N. et al. (2020). Toward a Model for Marking up Non-SI Units and Measurements. Journal of Text Encoding Initiatives. https://doi.org/10.4000/jtei.1996 (accessed 15 April 2022).

Monk, L. J. (1912). A Text Book of Stationery Binding. Raithby, Lawrence, and Co.

Pickwoad, N. (2009). In Suarez, M. S. and Turner, M. L. (eds), The Cambridge History of the Book in Britain. United Kingdom: Cambridge University Press. pp. 268-290.

Tomasek, K. and Bauman, S. (2014). Encoding Financial Records. http://journalofdigitalhumanities.org/author/ktomasek/ (accessed 15 April 2022).

Interactive Visualization of Indigenous Territory: Mapping of the Attikamegues in Historical Maps

Herold, Nastasia
nastasia.herold@uni-leipzig.de
Leipzig University, Germany

Blicher Christensen, Mathilde
blicher1997@gmail.com
University of Southern Denmark, Denmark

Jänicke, Stefan
stjaenicke@imada.sdu.dk
University of Southern Denmark, Denmark

The ancestral territory of the Indigenous nation of the Atikamekw is based in the current Canada. The nation claims its ancestral rights on the territory, but has been thwarted over the time by some historians doubting the Atikamekw being the descendants of the Attikamegues who were mentioned in historical reports of the first missionaries in New France (Dawson, 2003; Ratelle, 1987). The reason for the doubt is an ethnonymical change (ethnonym = name of a people) around the year 1700 - from "Attikamegues" to "Têtes-de-Boule". The so-called Têtes-de-Boule indicate their presence on the territory for milleniums, which is why they rechanged their official name to Attikamegues, now in their own orthography "Atikamekw" (http://www.manawan.org/accueil/).Some historians, above all Dawson, however, try to prove that the change around 1700 was not just an ethnonymical one, but an ethnical one, and that the Têtes-de-Boule are not the descendants of but a different people from the Attikamegues.

The humanist Nastasia Herold is currently writing her dissertation about the Atikamekw history. She involves the Atikamekw perspective in the scientific discussion and in addition to this, she includes data from the Eurocentric perspective that has not been considered yet. One of the Eurocentric data are English and French historical maps from much later than 1700 (until 1824 according to Hudson's Bay Company Archives, G.3/135) that mention a people called Attikamegues. This contradicts the hypothesis that the Attikamegues disappeared around 1700 and that a new people moved to their territory, the Têtes-de-Boule. These maps also illustrate that historical reports from missionaries cannot be used as the single source for a discussion on the history of an Indigenous nation.

Nevertheless, the mere existence of these maps does not prove the continuing existence of the Attikamegues/Atikamekw Nation. The newer maps could be copies of older maps, and/or the sources for the newer maps could be old reports from before 1700. In order to examine this consideration, metadata needs to be collected and compared, and the very different orthographies for "Attikamegues" should be checked against orthographies in reports in order to find the potential source of the maps.

We developed a Web-based framework to support investigating the above described research interest with visual interfaces. In order to do so, the following tasks are supported:

- **Collecting historical maps:** The framework allows for uploading historical map images and entering descriptive metadata about a map like publication year, publisher, etc. This information is necessary for all frontend components of the tool.
- **Annotating a contemporary map:** When uploading a historical map, the humanities scholar has to annotate a contemporary map with the approximate location of the ethnonym on the historical map. The assignment of a location to a map is a human decision, and domain knowledge as well as a comparison of the historical and contemporary topologies of the maps are essential. Related map annotation projects are maphub (https://maphub.github.io/) and Map Warper (https://mapwarper.net/).
- **Geospatial-temporal analysis:** The core element of the frontend are two linked views (see Figure 1): (1) a timeline displaying the distribution of when historical maps have been published, and (2) a map that illustrates a contemporary approximation of the historical Atikamekw territory. The ethnonyms are visualized in both views to support generating hypotheses regarding space, resembling a tag map (Reckziegel et al., 2018), and time. When hovering a dot in one view, the other view is automatically navigated to the particular time

period or geographical area. Clicking a dot shows all information on the chosen historical map.

Task-based filtering: The visual frontend is embedded in a faceted browser that allows for filtering particular sets of historical maps, for example by publication year to discover trends or by ethnonym type(s) for analyzing geospatial distributions.



Linked timeline and map to support explorative analysis of Atikamekw territory

The publication date in the timeline, the localization of the people corresponding to the historical map in the interactive map, and the original ethnonym both in timeline and interactive map give the researcher a better understanding of the relationship of these data, what makes it easier to evaluate if the historical map is a copy of an older map or not. The resulting hypothesis of copy "yes / no" can be saved in the metadata, and thereupon be filtered. The visualized result gives an overview of all maps that are potentially first editions. This is just one example besides many other hypotheses that can be developed using the interactive visualization.

The project was a perfect Digital Humanities collaboration: the research question, data collection, and desires of visualization and interactive tools were provided by a humanist, the design and implementation by computer scientists. The collaboration's development was discussed and progressed in weekly video conferences to make sure that the tools were helpful and useful for the humanist researcher. Furthermore, the project was designed in a way that even after the collaboration, maps and metadata can be added and modified by the humanist researcher. The project is published at a website (https://atikamekwstudies.org/historical-maps-live-tool/)for the benefit of Atikamekw community members and of researchers. The audience can even contribute by sending relevant historical maps to Herold that have not been included yet. This possibility of ongoing work and collaboration is an example of sustainable research in the Digital Humanities.

The "Atikamekw Historical Maps" project can be applied to other minorities all over the world. Cartographers of the last centuries were ordered to design maps of Asian regions for example. Very often, the localization of the peoples on the Asian islands and mainland was included in the maps for reasons of trade, politics, or military. Indigenous peoples in the Asian regions can adopt the Digital Humanities project described here for their needs in order to prove their continuous existence in a specific area.

The presentation is aimed at humanists, computer scientists, community members, and anyone else with an interest in collaborative DH research projects, mapping, visualization, and sustainable research in minority studies.

Bibliography

Dawson, N.-M. (2003). *Des Attikamègues aux Têtes-de-Boule.* Mutation ethnique dans le Haut Mauricien soule régime français. Sillery: Septention.

Jänicke, S. (2016). *Valuable research for visualization and digital humanities: A balancing act.* Workshop on Visualization for the Digital Humanities, IEEE VIS.

Ratelle, M. (1987). *Contexte historique de la localisation des Attikameks et des Montagnais de 1760 à nos jours.* Québec: Gouvernement du Québec.

Reckziegel, M., Cheema, M. F., Scheuermann, G. and Jänicke, S. (2018). *Predominance tag maps.* IEEE transactions on visualization and computer graphics, 24(6), 1893-1904.

Using Word Embeddings for Validation and Enhancement of Spatial Entity Lists

Herrmann, J. Berenike
berenike.herrmann@unibas.ch
University of Bielefeld, Germany

Byszuk, Joanna
joanna.byszuk@ijp.pan.pl
Institute of Polish Language, Polish Academy of Sciences, Kraków, Poland

Grisot, Giulia
giulia.grisot@uni-bielefeld.de
University of Bielefeld, Germany

Introduction

Spatial distant reading uses computational means to investigate fictional representations of space as a central category of sense-making (Lefebre, 1974), both in fictional

world building (e.g., Bologna, 2020) and in societal contexts (e.g., Wilkens, 2021).

Our spatial distant reading project investigates the affective topologies of German-Swiss literature to examine different types of spatial representation in fictional Swiss-German prose between 1854 and 1930, assessing iconic differences such as culture/nature, urban/rural (Rehm, 2014), as well as the (alpine) mountains' role in Swiss literary national framing (Zimmer, 1998). A key resource is a list of spatial terms (N=187,421 entities), including spatial named entities, other urban and rural toponyms, as well as natural terms (Grisot & Herrmann, in prep.). In the current paper, we take a methodological focus on this resource, exploring *word embeddings* for validation (Task A) and extension (Task B) of our spatial entity lists.

Word embeddings are a representation of words in the form of vectors in space, which describes their inter-relations within some language system in a spatial manner. Vectoral word representations go back to Wittgenstein's observation that word meaning is determined by the way they are used (Wittgenstein 1953: 80, 109), and Firth's "You shall know a word by the company it keeps." (Firth 1962: 11). Popularity of word embedding applications started by improved search engines (Mikolov et al. 2013), and is now used widely in humanities (Hamilton et al. 2016, Antoniak & Mimno, 2018). We use word embeddings to (a) validate our spatial entity detection lists compiled from external resources on word-vector representations; (b) develop a new resource for interior items that could not be obtained from external resources.

We used a large corpus comprising altogether N= 17,228 texts in German across different literary genres for the whole period of newer German literature (retrieved mainly from https://www.projekt-gutenberg.org/). Compilation ensured a maximally broad range of spatial terms, across time and genres, generalizing over the whole population of spatial terms in German (literature) with extensive training material, counting around N= 500 million tokens. Part of the corpus was our collection of German-Swiss literature, built using a combined list of 10,000 titles of fictional narratives written by Swiss authors in German (Herrmann et al., 2021), and a list of 1,997 names of authors with Swiss nationality, extracted from Wikidata. Open repositories were searched for digital versions of texts written by Swiss authors. A corpus of N= 482 Swiss novels was thus compiled, combining n= 450 novels in digital form, and n= 38 more novels newly digitized.

Method

We followed Ehrmanntraut et al.'s (2021) suggestion of fastText as the better performing solution for word embeddings in German. As our goal was to assess the usefulness of this method for discovery of terms related to terms associated with fictional space, with special attention on spatial interiors, we used a combined corpus of German and Swiss German literature to build a fastText model of vector relationships. We first built a model of spatial relationships and similarities as evidenced in German literature (extrapolated from the large and specialized collections described above). This model was examined using lists of seed words, requesting the 10 closest neighbors for each of the terms earlier distinguished as matching relevant categories by philological expertise. Our work was divided into two tasks with separate goals: Validating spatial entities lists (task A), Extending a manual list (task B).

Task A: Validating spatial entities lists

Goal of this task was validation of two entity lists compiled in a mixture of manual and automatic retrieval. We searched for the 10 most "similar" words for each item from the urban and rural spatial entities lists (for more details see Herrmann & Grisot, 2022):

- Rural entities (n= 274): spatial terms relating to – or characteristic of – the countryside, in particular human settlements or infrastructures, as opposed to those of the city (for example *Wanderweg*: footpath, *Feld*: field, *Hütte*: hut, shack);
- Urban entities (n= 262): spatial terms relating to the city, its buildings and infrastructures (for example *Bahnhof*: station, *Kreuzung*: crossing, *Palast*: palace)

For both lists we determined whether seed words from the same list appear among the similarity matches. High overlap indicates well-formation of the list, approaching the true population of spatial terms used in literature written in German for "urban" and "rural", respectively. We also perused the list for new relevant spatial entities to extend the list.

Task B: Extending a manual list

During our research, we noted that elements for the description of "interior" space were still missing from our spatial entities lists. Given that interiors are a vital part of not only realist literary narration (Fludernik, 2014), a resource was needed. In absence of a validated external resource, we use word embeddings to extend a small list of interior items. In the first step, a student assistant was instructed to compile a first list of objects and elements of furniture, but also of elements that are structural and yet part of a building/home environment (like 'table' or 'ceiling', but also 'column'). They were told to find as many terms as possible, from world knowledge, as well as by synonym lists (openthesaurus.de). The embeddings were used to discover new terms, with questions about which words were crucial for detecting the largest amount of other terms.

Results & Discussion

Task A rendered only little overlap among the ten most similar matches. Rather, we observed a high frequency of hyperonyms or specific variants of the seedwords, such as "Gartenpavillon", "Glas-Pavillon", "Teepavillon" (for "Pavillon", urban list), or synonyms such as "Trambahn", "Pferdebahn", "Autobus", "Straßenbahn", "Tramway", "Tramwaywagen", "Pferderlbahn", "Taxameterdroschke" (for "Tram", urban list). The same rich pattern was observed for Task B, the interiors list, which could indeed be appended by many elements (see repository). However, for the ensuing detection of spatial entities in our "affective spatial analysis," other procedures for parsing nouns into their components to increase term matching need to be considered as well, including for diminutives ("Dachstübchen", "Kartoffeläckerchen") and other within-word combinations. On a philological level, the observed degree of lexical specialization (e.g., "Gummibadewanne", rubber bath tub; interiors list) is interesting with respect to a realism effect (sensu Roland Barthes).

Altogether, using word embeddings provided a very constructive basis for improvement of our spatial entities lists, both for existing resources and the compilation of a new list. Our next step is to explore the distribution of the words obtained by the large language model in an exploratory analysis on our specialized corpus of German-Swiss literature around 1900.

Acknowledgments

The research was done as part of the Short Term Scientific Mission of JB at the Bielefeld University, funded by the Polish National Science Center (NCN) and the COST Action "Distant Reading for European Literary History (CA16204), supported by COST (European Cooperation in Science and Technology, www.cost.eu). JB was also supported by the Large-Scale Text Analysis and Methodological Foundations of Computational Stylistics project (SONATA-BIS 2017/26/E/HS2/01019). BH and GG were funded by "High Mountains Low Arousal? Distant Reading Topographies of Sentiment in German Swiss Novels in the early 20th Century" (SNSF)-COST-Project.

How big can a static site be? Staticizing a census database

Holmes, Martin
mholmes@uvic.ca
University of Victoria, Canada

Newton, Greg
gregster@uvic.ca
University of Victoria, Canada

Project Endings and static websites

It has long been recognized that building DH web applications which do not have perpetual funding on complex computing stacks that require regular updates presents an overwhelming problem for long-term maintenance and archivability (Nowviskie and Porter 2010; Dombrowski 2019; Smithies et al. 2019). Project Endings is a collaboration between DH scholars, librarians and programmers, aiming to create tools and recommendations for building extremely low-maintenance, easily archivable, but still highly functional digital edition projects (Goddard 2018; Holmes and Takeda 2019a). Starting in 2016, the Endings team have converted a number of high-traffic, well-known digital edition projects that previously ran on XML databases and similar back-end infrastructure into entirely static websites [1] built with HTML5, CSS and client-side JavaScript (Holmes 2017; Arneil, Holmes and Newton 2019). We have also developed a set of Principles [2] to guide us in converting existing projects and creating new ones, as well as a client-side pure-JavaScript search engine, staticSearch [3] (Holmes and Takeda 2020a, 2020b).

At the outset, we assumed that only relatively small digital editions would be suitable candidates for complete staticization, and we began with sites consisting of only a few hundred pages. [4] However, seeing how smoothly the process worked with smaller sites, we took on some of our larger projects, including The Map of Early Modern London (13,086 pages), The Colonial Despatches (10,826 pages), and Digital Victorian Periodical Poetry (20,685 pages). The results were very encouraging: the new sites were faster and more responsive than the old, and the staticSearch engine performed very effectively even at these scales. [5]

But there must, presumably, be some practical limits on the scale of a DH project which can be effectively converted to the static model, and it would be helpful to know where those limits might be. We reviewed our own catalogue and discovered a candidate which appears to be precisely the kind of project that would be suitable for testing this: *VIHistory*.

The *VIHistory* project

VIHistory (https://vihistory.uvic.ca/), a PostgreSQL/PHP project created about 15 years ago, presents census data from the City of Victoria and Vancouver Island, from censuses taken in 1871, 1881, 1891, 1892, 1901, and 1911, comprising nearly 150,000 individual census records, along

with associated tables of occupations, familial relationships, locations, addresses, religion, languages, nationalities and other associated concepts. In terms of the number of individual HTML pages required for a static site, it is between five and ten times the size of the largest project we have previously staticized. VIHistory has undergone successive infrastructure migrations over the years, and various features have broken as a result. A looming server migration will render it non-functional, so we must take action within a short period. Furthermore, a new dataset (an addendum to the 1901 census) is now available for addition to the collection. Rather than invest more time in patching the existing site, we are instead creating a static version, with a completion deadline of April 2022.

Census data brings with it a range of challenges, particularly when multiple censuses are to be presented as an integrated dataset. From one census to another, the range of data collected will vary; ward boundaries change; and descriptors such as nationality, race, occupation, familial relationships and languages mutate as social and political norms evolve. [6] The old version of the VIHistory site addressed these issues through a set of PostgreSQL views, which merged disparate datasets into a normalized form which could be queried more easily. One of our challenges will be to accomplish this in static form.

Another challenge will be to devise an appropriate granularity for the HTML pages which constitute the site. We expect to create a single HTML page for every distinct census entry, but other pages will be constructed to bring together collections of similar features (people with the same occupation, on the same street, with the same nationality, etc.) in order to provide a useful browsing approach to the data (something lacking in the existing site, which has only a search interface). The search itself will need to be carefully constructed so that all the features of the existing site search are preserved, and we expect to add more search options too.

Plan, approach, and prospects

We initially considered converting all the existing census data into XML as the first phase of the project, but we have not found any XML standard suitable for this historical census data. We then considered converting the data directly into HTML5, but found it more effective to design an intermediate custom XML schema, which enables records from all the different censuses to be encoded in a single flexible structure, and also permits us to apply datatype constraints and catch errors. The XML is then converted into XHTML5 for the website.

- Each census entry page will present a standardized tabular view of the data from the census record, with explanations to clarify differences between census datasets.
- Each "page" will have a condensed single-line title capturing essential data, used for display in search results and listings pages.
- Values for datapoints such as nationality, language and so on will be constrained by the schema, and all pages will be validated during the build process.
- Detailed diagnostics (Holmes and Takeda 2019b) will expose inconsistencies and errors in the dataset.
- Metadata will be encoded in the HTML header and used to create search filters, but the text of the pages will also be indexed for a new full-text search feature (the old site search allows only metadata filters).

We fully expect the conversion to succeed, but we also know that we will be pushing the practical limits of this approach, and will need to devise optimizations and workarounds to make the site, and particularly the static search engine, usable. Our paper will report on this work, and present recommendations on strategies and limitations for creating static resources on this scale.

Bibliography

Arneil, Stewart, Martin Holmes, and Greg Newton. 2019. "Clearing the Air for Maintenance and Repair: Strategies, Experiences, Full Disclosure; Paper Three: Ruthless Principles for Digital Longevity." Presented at Digital Humanities 2019, Utrecht, the Netherlands. https://dev.clariah.nl/files/dh2019/boa/0648.html.

Dombrowski, Quinn. 2019. "Sorry for all the Drupal: Reflections on the 3rd anniversary of 'Drupal for Humanists.'" Quinn Dombrowski (blog), November 8, 2019. http://www.quinndombrowski.com/?q=blog/2019/11/08/sorry-all-drupal-reflections-3rd-anniversary-drupal-humanists.

Goddard, Lisa. 2018. "The Endings Project @ UVic: Concluding, Archiving, and Preserving Digital Projects for Long-Term Usability." @Risk North 2: Digital Collections, Montreal, Canada. https://github.com/projectEndings/Endings/blob/master/presentations/Goddard_RiskNorth_Endings_final.pptx?raw=true.

Holmes, Martin. 2017. "Selecting Technologies for Long-Term Survival." Presented at the SHARP Conference 2017: Technologies of the Book, Victoria, BC, Canada. https://github.com/projectEndings/Endings/raw/master/presentations/SHARP_2017/mdh_sharp_2017.pdf.

Holmes, Martin. 2021. "Using ODD for HTML." *The Journal of the Text Encoding Initiative.* Text Encoding

Initiative Consortium. https://journals.openedition.org/jtei/3106.

Holmes, Martin and Joseph Takeda. 2019a. "The Prefabricated Website: Who needs a server anyway?" Text Encoding Initiative Conference, Graz, Austria. https://zenodo.org/record/3449197.

Holmes, Martin and Joseph Takeda. 2019b. "Beyond Validation: Using Programmed Diagnostics to Learn About, Monitor, and Successfully Complete Your DH Project." *Digital Scholarship in the Humanities.* Oxford University Press/EADH. http://dx.doi.org/10.1093/llc/fqz011.

Holmes, Martin, and Joey Takeda. 2020a. "Static Search: An Archivable and Sustainable Search Engine for the Digital Humanities." Presented at the Digital Humanities Summer Institute (DHSI) Colloquium (#VirtualDHSI). [https://zenodo.org/record/3883150].

Holmes, Martin, and Joey Takeda. 2020b. "Nine Projects, One Codebase: A Static Search Engine for Digital Editions." Presented at the COLLABORATION Digital Humanities Conference, University of British Columbia / online. http://dhconference.sites.olt.ubc.ca/conference-info/program/day-4/.

Nowviskie, Bethany, and Dot Porter. 2010. "Graceful Degradation: Results of the Survey." Presented at Digital Humanities 2010, King's College, London. https://nowviskie.org/Graceful_Degradation.pdf.

Smithies, James, Carina Westling, Anna-Maria Sichani, Pam Mellen, and Arianna Ciula. 2019. "Managing 100 Digital Humanities Projects: Digital Scholarship & Archiving in King's Digital Lab." *Digital Humanities Quarterly* 13, no 1 (2019). http://www.digitalhumanities.org//dhq/vol/13/1/000411/000411.html.

Stanger-Ross, Jordan. 2008. "Citystats and the History of Community and Segregation in Post-WWII Urban Canada." *Journal of the Canadian Historical Association* 19, 2 (2008), 3-22. https://citystats.uvic.ca/Citystats_CHA_19.2.pdf.

Notes

1. See https://endings.uvic.ca/projects.html for the full list of staticized projects.
2. https://endings.uvic.ca/principles.html.
3. https://github.com/projectEndings/staticSearch.
4. See for example *My Norse Digital Image Repository* (https://myndir.uvic.ca/) or (*The Robert Graves Diary* (https://graves.uvic.ca/).
5. See https://mapoflondon.uvic.ca/search.htm; https://bcgenesis.uvic.ca/search.html; https://dvpp.uvic.ca/search.html.
6. See, for example, Stanger-Ross 2008, which discusses the evolution of ethnicity in census data.

PRISMS: a new platform for digital Book History

Huber, Alexander
alexander@hubers.org.uk
Huber Digital

Huber, Emma
emma@hubers.org.uk
Huber Digital

Introduction and motivation

In a 2020 talk entitled "A Hornbook for Digital Book History", Whitney Trettien weaves together many of the strands that have led book history, bibliography, media studies, and the digital humanities to have become deeply entangled in recent years. She convincingly argues for the potential of Book History done digitally "to build connective tissue across scattered collections" and advocates "using the digital tools at our disposal in order to see the big picture of the past". [1]

It is in this vein that this paper presents the motivation for and realisation of a new open-access open scholarship platform (currently in public beta) named PRISMS. [2] The aim of the PRISMS Open Scholarship platform is two-fold:

1. It offers a publication platform for digital scholarly editions, with full-text (preferably encoded in TEI) and facsimiles, and any accompanying materials, such as introduction, editorial statement, critical apparatus, contextual source materials, bibliography, and indices;
2. It facilitates the semantic annotation of these editions and their related scholarship (in any format) by enabling easy-to-perform formal ontological modelling (based on the CIDOC-CRM family of ontologies [3]), and thus hopes to contribute to providing the abovementioned "connective tissue" not only for scattered collections, but to overcome the artificial print/digital divide.

PRISMS was born out of the realization that digital editions do not break with the historicity or materiality of the sources they organize and present, but instead remediate and extend them in ways that enable new forms of access, engagement, presentation, and analysis. PRISMS

conceptualizes digital editions as living entities that perform rather than merely document the remediation they engage in.

The scholarship that underpins each digital edition provides the essential context for these remediation processes, and collectively they sustain the knowledge network that supports all academic engagement with the texts from any disciplinary viewpoint. PRISMS is designed to allow for the collaborative and collective modelling of this continuum of digital editions and scholarship by placing digital editions, their material and contextual basis, and the resulting academic engagement in a linked context, building on the standards and tools provided by the Semantic Web.

We believe that this type of formalization is beneficial for the purposes of this project in at least three ways: firstly, ontologies facilitate modelling with reduced reliance on implicit knowledge through an explicit, shared conceptualization of the domain. Secondly, formal models encourage collaboration as they can be shared, re-used, adapted (forked), enhanced, aggregated, and developed collaboratively. Thirdly, as a form of knowledge representation, visualization, and preservation, formal models support computational processing and ultimately reasoning, and can develop alongside the mental models and human reasoning we engage in as scholars. PRISMS facilitates scholarship that is based on these principles [4].

Approach and implementation

The PRISMS Open Scholarship platform integrates the task of publishing digital editions with the need for analytical and modelling tools to perform the type of knowledge representation that connects the material, digital, and the scholarship that builds on them. PRISMS aims to support digital editors, book historians, experts in media and cultural studies, librarians, literary scholars, and of course digital humanists, to ensure a wide range of domain expertise, disciplinary practices, and methodological approaches are reflected in the platform. To this end, the PRISMS platform hosts a variety of tools alongside the digital editions, which can be categorized as component tools (such as text-based tools, image-based tools, XML-based tools, etc.) and workbench tools (those available across document types and editions).

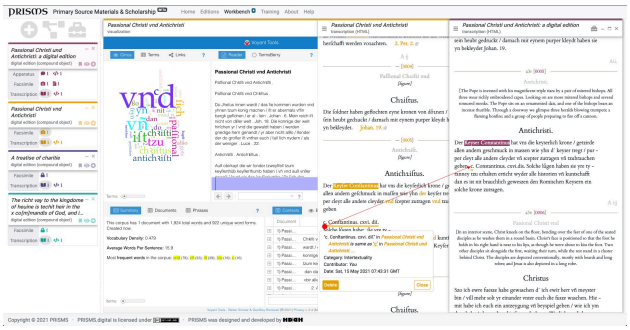


Figure 1
Some of the built-in analysis and visualization tools in PRISMS. Voyant-Tools is shown alongside a relation being made between two editions, and some highlighted annotations

The former category of tools is useful for any type of close scholarly work, and in PRISMS these tools include a bookmarking tool, an annotation tool (initially focussing on texts and images, but with a vision to extend annotation capabilities across all media types), the ability to keep research records (and other forms of note-taking, e.g. transcriptions, translations etc.) in the form of notebooks, and integration of Voyant Tools [5] for statistical analysis and a variety of visualisations of texts. The latter category includes the ability to participate in the shaping of the knowledge graph by modelling concepts and relationships, an easy way to organize research materials, and the ability to download, share, and publish contributions for the benefit of all.

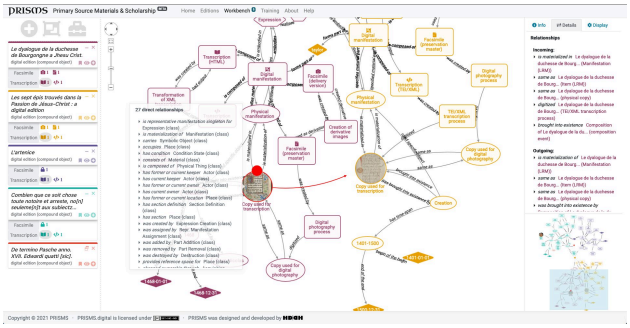


Figure 2
Modelling both the material legacies and digital remediation processes with Linked Data technologies and Cytoscape.js

All semantic modelling work, e.g. with regard to the provenance of the material and digital manifestations of an edition, can be performed both directly in a visual representation of the graph using Cytoscape.js or via a set of customizable HTML forms. All resulting triples are stored in an RDF-native graph database (using the

abovementioned ontologies) for long-term preservation, collaboration, and re-use. Every user of PRISMS has both read access to this global graph that underpins the platform and unlimited access (via SPARQL Update operations) to a private graph. By default, everything in PRISMS is private. Everything contributors add is immediately visible to them, and they can conduct their scholarship in complete privacy, without delay or interference. Only when the user decides to publish their contributions will they be made available to everyone. Depending on the type of contribution made, there are options to share, download, and/or publish them. All contributions made to the PRISMS platform are stored in standard formats, e.g. the W3C Web Annotation Data Model for annotations and relations.

Contribution and further work

PRISMS has been launched with corpora from the EEBO-TCP 6, ECCO-TCP 7, EVANS-TCP 8, the DTA extended core corpus 9, and the Taylor Editions 10 scholarly editions platform. And it is easy to add new editions to the platform either as part of a dedicated digital scholarly editing process, for which training is provided, or by simply adding a IIIF manifest and using a built-in XML-aware or standard text-editor to start transcribing and adding contextual materials. The platform already supports the addition and semantic annotation of a wide range of primary and secondary materials, such as facsimiles in IIIF, a transcription of a source text, a PDF of a journal article, a video of a theatrical performance, an audio book, an image, or a 3D-model of a sculpture mentioned in a text, etc.

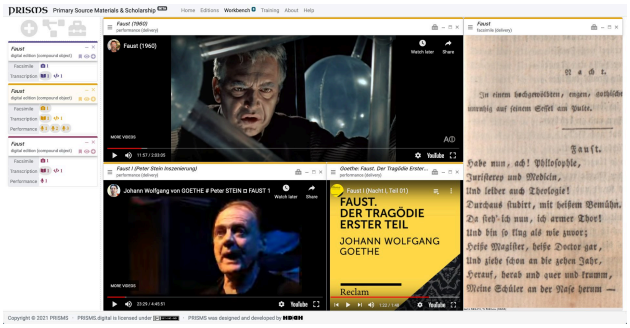


Figure 3
A view of the PRISMS workbench, with three editions of Faust loaded, and an aggregation of primary and research materials in support of a performance analysis, with a facsimile, two videos, and an audio book

Moving forward, we will continue to work on integrating the digital research and tools important to PRISMS' users (e.g. reference manager, images taken in reading rooms, items deposited in institutional repositories). With end of the beta phase, the project also intends to provide access to the entire PRISMS knowledge graph through regular data dumps and a public SPARQL endpoint. We envision that over time, PRISMS will evolve both as a powerful discovery tool and a personal research tool.

Bibliography

Ciotti, F. (2015) "Digital methods for Literary Criticism." Lecture slides. University of Rome Tor Vergata, http://didattica.uniroma2.it/files/scarica/insegnamento/161783-Informatica-Umanistica-Lm-Per-Il-Llea/37175-Slide

Ciula, A. and Marras, C. (2016) "Circling around texts and language: towards 'pragmatic modelling' in Digital Humanities." *Digital Humanities Quarterly (DHQ)* 10.3 http://www.digitalhumanities.org/dhq/vol/10/3/000258/000258.html

Ciula, A. and Eide, Ø. (2107) "Modelling in digital humanities: Signs in context." *Digital Scholarship in the Humanities* 32: i33–i46. https://doi.org/10.1093/llc/fqw045

Ciula, A., Eide, Ø, Marras, C. and Sahle, P. (2018) *Models and Modelling between Digital and Humanities — A Multidisciplinary Perspective.* Historical Social Research (HSR) Supplement 31.

Eide, Ø. (2015) *Media Boundaries and Conceptual Modelling: Between Texts and Maps.* Pre-print manuscript, https://www.oeide.no/research/eideBetween.pdf

Kirschenbaum, M. and Werner, S. (2014) "Digital Scholarship and Digital Studies: The State of the Discipline." *Book History* 17, 406-458 https://www.academia.edu/15995371/Digital_Studies_and_Digital_Scholarship_The_State_of_the_Discipline

Kräutli, F. and Valleriani, M. (2018) "CorpusTracer: A CIDOC database for tracing knowledge networks." *Digital Scholarship in the Humanities* 33(2): 336-346. https://pure.mpg.de/rest/items/item_2472866_10/component/file_3002633/content

Laehnemann, H. (2022) "History of the Book blog." https://historyofthebook.mml.ox.ac.uk/

Oldman, D., Doerr, M. and Gradmann, S. (2016) "Zen and the Art of Linked Data: New Strategies for a Semantic Web of Humanist Knowledge." In Schreibman, S., Siemens, R., Unsworth, J. (eds.) *A New Companion to Digital Humanities.* Malden, MA: Wiley Blackwell, 251-273.

Notes

1. https://rarebookschool.org/rbs-online/a-hornbook-for-digital-book-history/. She shares this vision of a continuum of print and digital with other influential

voices at the intersection of book history, media studies, and the digital humanities, among them Henrike Laehnemann, Sarah Werner, and Matt Kirschenbaum to name but a few.
2. https://www.prisms.digital/
3. http://www.cidoc-crm.org/
4. Ground-breaking research projects in this domain include the ResearchSpace platform and the Sphaera CorpusTracer project.
5. https://voyant-tools.org/
6. https://textcreationpartnership.org/tcp-texts/eebo-tcp-early-english-books-online/
7. https://textcreationpartnership.org/tcp-texts/ecco-tcp-eighteenth-century-collections-online/
8. https://textcreationpartnership.org/tcp-texts/evans-tcp-evans-early-american-imprints/
9. https://www.deutschestextarchiv.de/
10. https://editions.mml.ox.ac.uk/

Everyday memory: A computational analysis of changing relation between past and present in Dutch newspapers in the twentieth Century

Huijnen, Pim
p.huijnen@uu.nl
Utrecht University, Netherlands, The

Historians have since long stressed the political and cultural functions of memory and heritage for societies. The more born-digital data have a past of their own, the more memory has also become a topic of interest for data scientists (Au Yeung and Jatowt 2011, Keegan and Brubaker 2015, Graus et.al. 2018, West and Leskovec 2021). Stringently connected to memory but much less studied in data science is our experience of time (Jatowt et.al. 2015 and Van Eijnatten and Huijnen 2021 being notable exceptions). The influential theory of François Hartog, for example, explains the memory boom from the establishment of a 'time regime of the present' at the end of the 1980s (Hartog 2015). Aleida Assmann goes against Hartog's subsequent assertion that an increasingly shorter present is all we are left with (Assmann 2020, 139). Instead, she puts forward her notion of cultural memory to replace ideas of temporal ruptures with a model that 'emphasizes the ineluctable entanglement of [past, present and future]' (Assmann 2020, 195-6).

Departing from this theory, this study proposes a computational approach to study the relation between the present and the past in twentieth century newspapers by analysing trends in phrases of the format 'n years ago'. Obviously, there is a plethora of ways in language can evoke the past. However, there are two arguments that justify singling out 'n years ago'. First, 'n years ago' is, as an expression, ubiquitous and syntactically stable in Dutch newspaper language throughout the studied period of the twentieth century. Second, unlike explicit references to years, events or persons in the past, 'n years ago' intricately ties the past to the present. The phrase presents the past—and keeps it present— *as something useful for the present*. This notion of 'useful past' or 'present past' (Paul 2015, 25-27) forms part of Assmann's critique of Hartog's theory of temporal orders.

Questions and method

The questions that are at the center of this study are how trends in references to present pasts relate to named philosophies of time experience, but also which past remains present and how these trends change over time. It takes newspapers as data, because of the vital role they as the 'first rough draft of history' have played in memory culture throughout the twentieth century. This study is based on the digitized versions of the most important nation-wide and regional newspapers the Dutch National Library holds and has made available. [1] The preliminary results presented here are based on the example of the national newspaper Telegraaf (1893-1989) of 10,000 documents (articles and advertisements) per year. [2] These documents have been rid of duplicates and cleaned with the help of Python's NLTK package. [3]

Analysis and results

The analysis is done with Python scripts in Jupyter Notebooks and consists of three subsequent steps:

The first step is the extraction of a list of the most common trigrams ending with 'year(s) ago' [4] from the cleaned and sampled dataset. This list is sorted by decade and by frequency to get an idea of the years that newspapers most often use in the phrase 'n years ago' throughout the twentieth century. This indicates that single digit years (one – nine) make up the most common phrases of 'n years ago', along with decades (ten, twenty, etc.) and one hundred.

Table 1: Most common trigrams ending in 'year(s) ago' with their English translation from a sample of 10,000 documents from the national newspaper De Telegraaf per decade per million trigrams, 1890-1980.

Trigram	English translation	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980
Een jaar geleden	One year ago	1,95	4,02	4,62	8,81	21,77	13,23	16,06	19,51	17,90	16,16
Twee jaar geleden	Two years ago	1,70	2,10	2,70	3,18	7,51	5,60	12,67	22,43	27,79	31,44
Twintig jaar geleden	Twenty years ago	1,27	0,50	0,59	2,34	3,28	4,24	3,39	3,34	4,07	6,08
Paar jaar geleden	Few years ago	1,10	1,26	1,06	1,50	1,53	3,39	2,74	7,05	9,22	10,01
Vier jaar geleden	Four years ago	0,76	0,17	0,53	1,20	2,48	1,36	3,39	7,37	10,85	11,99
Drie jaar geleden	Three years ago	0,68	1,26	1,06	1,86	3,54	2,88	6,22	11,08	14,50	14,75
Dertig jaar geleden	Thirty years ago	0,59	0,42	0,26	0,72	1,86	2,71	2,42	2,92	1,89	3,09
Tien jaar geleden	Ten years ago	0,51	1,01	0,79	1,80	4,16	4,24	5,25	9,54	12,39	15,42
Honderd jaar geleden	One hundred years ago	0,42	0,50	0,73	1,14	2,08	2,54	1,86	2,86	2,02	1,98
Acht jaar geleden	Eight years ago	0,42	0,25	0,20	0,60	0,91	0,85	1,05	2,97	3,20	3,59

In a second step, the most common references of 'n years ago' are plotted, relative to one another, over time. Figures 1-3 show the trajectories of all single years together (figure 1), of all decades from ten to one hundred (figure 2) and of the single years one, ten, one hundred and two hundred (figure 3). These figures show that Dutch newspapers started to look back in time by the use of the phrase 'n years ago' since the 1930s. Once they did, the use of some variations of this phrase strongly gained traction, particularly in reference to the near past (one to ten years ago).

The final step of the analysis is to look at the actual years that the phrase 'n years ago' referred to throughout the twentieth century. Do newspapers tend to look back at specific years, as in the end of the Second World War being 'five years ago' in 1950? Or are 'notable' years submerged in what is here called 'everyday memory culture'? The latter seems to be the case if these years are calculated for 'n years ago', where n stands for the years from one to twenty and decades from twenty to two hundred. Figure 5 shows that 'n years ago' tends to evoke the recent past itself above any particular year. No single year really stands out. [5]

Conclusion

In the light of Hartog's theory of an all-encompassing present, one would have expected a steady decrease of references to the past and the future. This study shows the opposite and substantiates Assmann's contention that interest for the past returns, particularly in the second half of the twentieth century, in the form of memory culture. With the growing popularity of the studied phrase, Dutch newspapers became mediators of a form of memory culture than can be seen as 'latent' or 'everyday' in that it emphasizes recurring events rather than returning to a specific thing in the past. This is, particularly, true for the phrases 'two years ago' and 'four years ago' (Figure 4), the spikes in the diagrams of which indicate references to important sport (Olympics, European and World Championship soccer) and political events (national parliamentary elections).

The use of the phrase 'n years ago' is by no means the only, nor the most important manifestation of memory culture. This limits the explanatory power of this study. In contrast with official and cultivated forms of memory culture, however, it does shed light on the latent, almost oblique, everyday invocation of the past that forms just as much part of that culture.

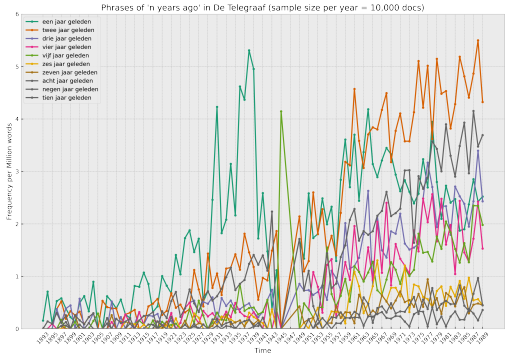


Figure 1:
Frequency over time of 'n years ago', where n stands for one to ten, per million words in a sample of 10,000 documents per year of the national newspaper De Telegraaf*, 1893-1989. For similar diagrams (figures 1-4) for other newspapers, see: https://github.com/PimHuijnen/looking_back_newspapers/tree/main/Data.*

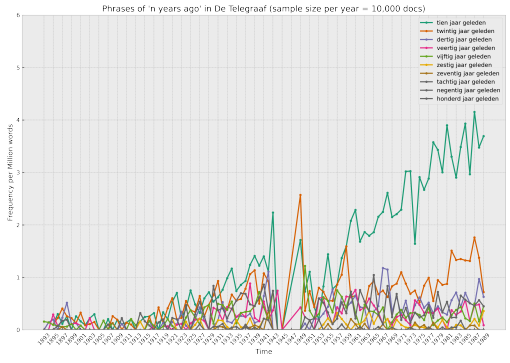


Figure 2:
Frequency over time of 'n years ago', where n stands for decades from ten to one hundred, per million words in a sample of 10,000 documents per year of the national newspaper De Telegraaf*, 1893-1989.*

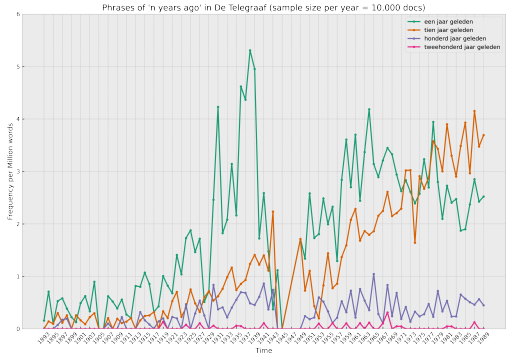


Figure 3:
Frequency over time of 'n years ago', where n stands for one, ten, one hundred and two hundred, per million words in a sample of 10,000 documents per year of the national newspaper De Telegraaf*, 1893-1989.*

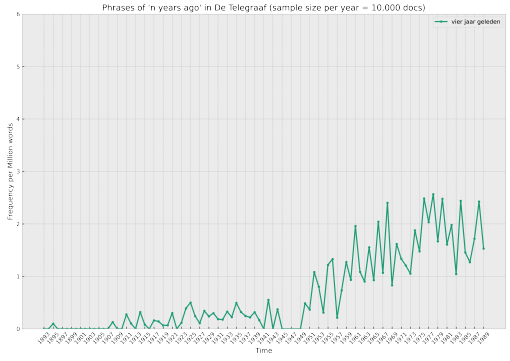


Figure 4:
Frequency over time of 'four years ago' per million words in a sample of 10,000 documents per year of the national newspaper De Telegraaf*, 1893-1989.*

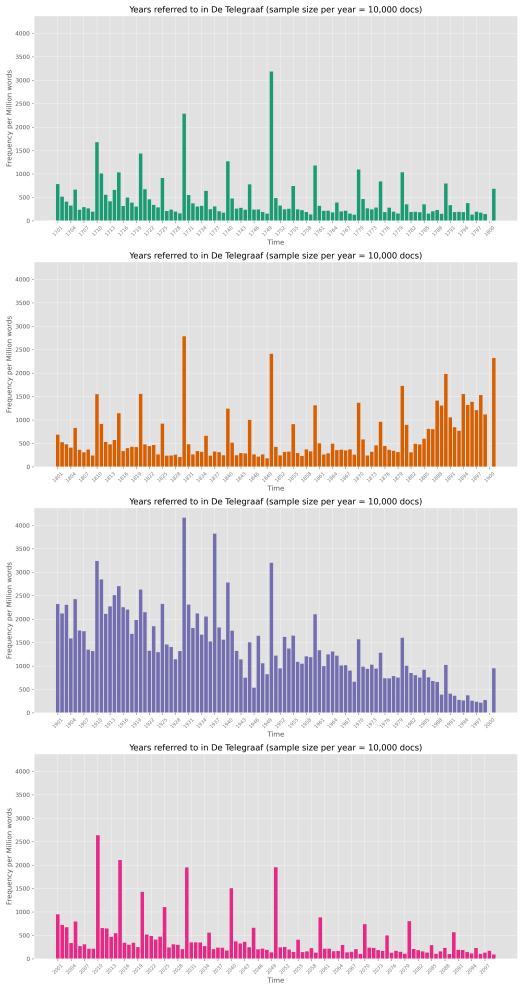


Figure 5:
Frequency over time of dates (in years) that are referred to via the phrase 'n years ago', where n stands for years from one to twenty and decades from twenty to two hundred, per million words in a sample of 10,000 documents per year of the national newspaper De Telegraaf*, 1893-1989.*

Bibliography

Assmann, Aleida (2020). *Is Time out of Joint? On the Rise and Fall of the Modern Time Regime.* Ithaca: Cornell University Press.

Au Yeung, Ching-man and Jatowt, Adam (2011). *Studying how the Past is Remembered: Towards Computational History through Lare Scale Text Mining. Proceedings of the 20 th ACM International Conference on Information and Knowledge Management*, 1231-1240. DOI: https://doi.org/10.1145/2063576.2063755.

Eijnatten, Joris van and Pim Huijnen (2021). Something Happened to the Future: Reconstructing Temporalities in Dutch Parliamentary Debate, 1814-2018. *Contributions to the History of Concepts*, 16: 52-82. DOI: https://doi.org/10.3167/choc.2021.160204.

Graus, David, Daan Odijk and Maarten de Rijke (2018). The Birth of Collective Memories: Analyzing Emerging Entities in Text Streams. *Journal of the Association for Information Science and Technology*, 69: 773-786. DOI: http://dx.doi.org/10.1002/asi.24004.

Hartog, François (2015). *Regimes of Historicity: Presentism and Experiences of Time.* New York: Columbia University Press.

Jatowt, Adam et.al. (2015). Mapping Temporal Horizons: Analysis of Collective Future and Past related Attention in Twitter. *WW '15: Proceedings of the 24 th International Conference on World Wide Web*, 484-494. DOI: http://dx.doi.org/10.1145/2736277.2741632.

Keegan, Brian C. and Jed R. Brubaker (2015). "Is" to "Was": Coordination and Commemoration in Posthumous Activity on Wikipedia Biographies. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing,* 533-546. DOI: https://doi.org/10.1145/2675133.2675238.

Paul, Herman (2015). *Key Issues in Historical Theory.* New York and London: Routledge.

West, Robert, Jure Leskovec and Christopher Potts (2021). Postmortem Memory of Public Figures in News and Social Media. *Proceedings of the National Academy of Science* 118. DOI: https://doi.org/10.1073/pnas.2106152118.

Notes

1. www.delpher.nl/kranten .
2. For the final paper, these titles will be complemented with other available newspaper titles with similar century-spanning scopes. Subsequently, all results will be aggregated to come to an encompassing picture of the present past in Dutch newspapers by means of references to 'n years ago'. The Telegraaf was chosen as an example here, because it shows trends that are paradigmatic for most other newspapers.
3. https://github.com/PimHuijnen/looking_back_newspapers . Cleaning in this script is mostly restricted to the removal of punctuation and caps. Stop words are not removed to guarantee that the Dutch word for the English indefinite article 'a' ('een'), which is part of any standard stop word list, remains part of the data. Similarly, the removal of numbers is no part of preprocessing to allow for phrases like '10 years ago'(even though Dutch linguistic convention formally does not allow for numbers in running text).
4. The Dutch equivalent of 'year(s) ago' is for both one and more years written in the singular form 'jaar geleden'.
5. Given the nature of this method, the decline in the frequency of years after 1975 that Figure 5 shows is as necessary as it is meaningless. There is, after all, increasingly less data (that ends in 1989) on which these numbers can be based.

Mining for clean energy: a machine learning approach to historicized sentiment mining of fossil fuel discourse in the Netherlands

Huijnen, Pim
p.huijnen@uu.nl
Utrecht University, Netherlands, The

Plets, Gertjan
g.f.j.plets@uu.nl
Utrecht University, Netherlands, The

Verheul, Jaap
j.verheul@uu.nl
Utrecht University, Netherlands, The

Global sustainability is one of the most urgent issues of today. Although the climate crisis has been on and off the Dutch political agenda for at least fifty years, a longer-term historical perspective on the crisis plays a minor role in present-day discussions. Nevertheless, such a perspective can give us important insights into the forces at play in the complex environmental, social, and cultural issues that are at stake when it comes to sustainability. We cannot explain

how and why public sentiments have changed over the last decades without mapping where they are rooted in and how they have evolved over time. We aim to investigate this for the Dutch case by tracing shifting sentiments towards fossil fuels in the postwar newspaper discourse. Interesting, for example, is that natural gas was framed as a sustainable, environment-friendly alternative for coal and petrol when it was introduced in the 1960s, while it presently carries the same deprecative label of 'fossil fuel' as the others. We have built a sentiment mining pipeline to be able to better understand semantic and emotional shifts like these.

Mass digitization has made a long-term historical semantic perspective on public meaning, emotions, and sentiments, as this project envisions, both innovative and feasible. The training of dedicated models based on the vectorization of language, at the same time, has enabled studying semantics on an entirely different scale than manually possible. The state-of-the-art in language models are transformer models, which, unlike previous vectorization techniques like word embeddings, consider the context in which words are used thanks to the introduction of "self-attention" layers. Therefore, they result in a more precise modeling of features of language than previous models, as Google's BERT has demonstrated (Devlin et.al. 2019).

This project deals with the analysis of sentiment within newspaper articles from 1960 to 1995 contained in the massive digitized newspaper archive of The National Library of the Netherlands (KB) [1]. By creating multiple fine-tuned BERT models, adapted to topics and decades, this project has produced models that run historically dynamic sentiment analyses that are context-specific and easily repeatable on different topics. The output of the models creates a sentiment variable which is topic- and context-specific.

Preprocessing and selection

The original OCR newspaper texts received from KB have been reformatted and divided by decade. Only texts labeled as articles (not advertisements) category were considered, which resulted in a final pre-training dataset of 43.4GB of uncompressed text. On this dataset, we have used a labeling and predicting pipeline to extract documents on fossil fuels, and a second to predict sentiments.

Before labeling, the data was first cleaned and tokenized. We used the SentencePiece (Kudo and Richardson, 2018) library to create a tokenizer file. Then we followed the lead of the Swedish (Malmsten et al., 2020), Finnish (Virtanen et al., 2019) and German (Branden et al., 2019) BERT project to select dictionary size and the configuration of the BERT model. Labeling the sentiment of articles is a complex task, as an article is composed of many sentences that might have contrasting sentiment when taken individually and/or out-of-context. The data was, further, decomposed in its main paragraphs as divided by KB's OCR to predict topicality and sentiment on a more fine-grained level. [2] Labeling for topicality was done per type of fossil fuel (coal, natural gas, petrol) and per decade by two labelers (which the project team evaluated), resulting in a final dataset of 1.5GB and 568,160 paragraphs of text with a 0.95 confidence score (see table 1). This data was used as input for the later models.

Table 1: Dataset after prediction of topicality: newspaper articles on goal, natural gas, petrol for the 1960s-1990s.

Decade	Type of fossil fuel	Size (MB)	No. paragraphs
1960s	Coal	27	4,626
	Natural Gas	102	40,816
	Petrol	172	57,196
	Total	301	102,638
1970s	Coal	15	5,388
	Natural Gas	114	51,678
	Petrol	245	96,189
	Total	374	153,255
1980s	Coal	79	29,289
	Natural Gas	389	174,378
	Petrol	7	1,474
	Total	475	205,141
1990s	Coal	11	2,259
	Natural Gas	61	16,127
	Petrol	384	88,740
	Total	456	107,126

Sentiment Labeling and fine-tuning

We selected two labelers to label the sentiment on each paragraph. This was done to improve the generalizability of the models and to avoid that the models would learn the subjective interpretation of one labeler on the articles' sentiments. The labelers used a range of three classes: -1 (negative), 0 (neutral), and +1 (positive). After labeling, we calculated the interrater reliability score using Cohen's Kappa score (Cohen, 1960) as a measure of the agreement among raters with the objective to compute the extent to which the data collected in the study are correct representations of the variables measured. The computed scores highlight a low agreement on 1960s and 1970s (average 0.22 and 0.24); a higher agreement on 1980s (0.36) and the highest agreement on 1990s (0.58). We weighted the labelers' judgements of the same paragraphs in the

following way: if they disagreed between 0 and -1/+1, we used the more 'extreme' label; paragraphs with opposite labels were discarded altogether.

Instead of the pre-training we opted to fine-tune BERT models, using an already pre-trained BERT model. The model selected to be fine-tuned with the sentiment labels and texts classified by our labelers was BERTje (de Vries et al., 2019). The fine-tuned BERT models take the labels for each type of fossil fuel (natural gas, coal, petroleum) within one decade (1960s – 1990s) to predict sentiment scores for the entire datasets.

Results

Based on the predictions, we have been able to visualize average sentiments (-1 to +1) per year for the entire newspaper discourse on the three types of fossil fuel between 1960 and 1995 (Figure 1). Most striking is the fact coal rather that natural gas is regarded more positively after the 1973 oil crisis (although the Dutch government had in 1965 already decided to stop all coal production by 1975). These trends will form the basis for an in-depth analysis in our final paper of the fossil fuel discourse in the Netherlands that will be based on significant changes in notable words (tf-idf) over time and between the three types of energy. We will, particularly, focus on shifts in discourses related to nuclear energy and renewable energy. Idf scores indicate that the former becomes increasingly noticeable in articles on natural gas throughout the decades, while the notion of 'clean' ('schone') becomes less prominent (Figure 2).

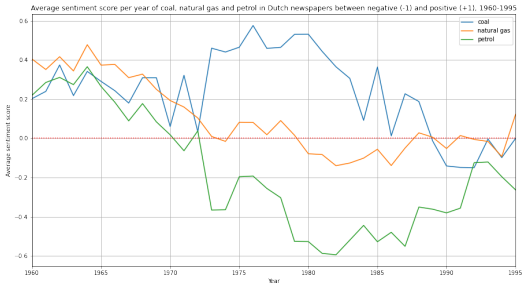


Figure 1:
Average sentiment score of articles on coal, natural gas, and petrol in Dutch newspapers between negative (-1) and positive (+1) per year, 1960-1995

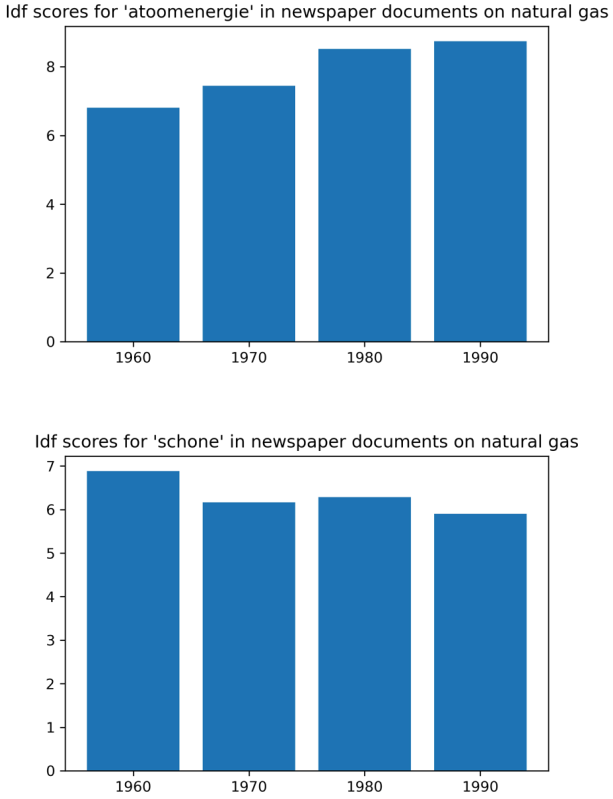


Figure 2:
Idf scores (per decade) for 'nuclear energy' ('atoomenergie') and 'clean' ('schone') in documents on natural gas, 1960s-1990s.

Bibliography

Branden C. et al. (2019). German's Next Language Model. *arXiv preprint arXiv:2010.10906.*

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.

Delobelle, P., Winters, T., & Berendt, B. (2020). Robbert: a dutch roberta-based language model. *arXiv preprint arXiv:2001.06286*.

De Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., & Nissim, M. (2019). Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

Kudo T. and Richardson J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. *arXiv preprint arXiv:1808.06226*.

Malmsten M. et al. (2020). Playing with Words at the National Library of Sweden – Making a Swedish BERT. *arXiv preprint arXiv:2007.01658.*

Virtanen A. et al. (2019). Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076.*

Notes

1. https://www.kb.nl.
2. We split paragraphs longer than 400 words and discarded very short (one-sentence) paragraphs. All preprocessing and modeling scripts are available on https://github.com/UtrechtUniversity/hist-aware.

Participatory Action Research for a Digital Humanities research project: Investigating Open GLAM in the context of Social Movement Archives

Humbel, Marco
marco.humbel.17@ucl.ac.uk
UCL Department of Information Studies, United Kingdom

Brief abstract

This paper responses to the question: How can the methodology of Participatory Action Research (PAR) be used to investigate Open Access to digital collections in the context of the Marx Memorial Library London (MML)?

Motivation

PAR is an established methodology in library-, archive-, and information studies for collaborating with practitioners, and members of the public in a research project (Pickard, 2013: 157–66). The potential of knowledge co-production through participatory frameworks receives also increasingly attention where DH research questions are investigated by means of qualitative data (Ortolja-Baird and Nyhan, 2021: 17–18). While PAR has been used for Digital Humanities (DH) projects (Pringle, 2020: 10–11; Ruge et al., 2016: 4–5), the methodology is however not present in recent DH method books (Levenberg et al., 2018; Schuster and Dunn, 2020). Through a case study of applying PAR in a PhD project, and a reflection on the research process with reference to the literature, this paper offers an introduction to the methodology and a set of transferable lessons-learned that could devise future PAR DH projects.

Research context

Open Access to digitized collections, also known as Open GLAM (Galleries, Libraries, Archives, and Museums), "[…] refers to a policy or practice that allows reuse and redistribution of materials for any purpose, including commercial" (Wallace, 2020a). However, a lack of resources and expertise hamper especially smaller institutions to digitize and to release collections as Open Access (Wallace, 2020b: 2–3). This project has used PAR to investigate Open GLAM in the context of archives with few resources, but which understand archiving as a form of activism to collect the histories of those who are marginalized in the historical canon: Social Movement Archives (Flinn, 2011; Hoyer and Almeida, 2021). The archive I have collaborated with is the MML, where I have volunteered from 2018 to 2021.

Applying PAR

PAR involves the stages of: identifying a desired change, planning an action, taking action, and evaluating on the action's outcomes. Based on the evaluation and reflection on the action, a new action may be enacted, which gives PAR a cyclical nature (Kemmis and McTaggart, 2005: 563).

I have identified PAR as an appropriate methodology because it allowed me to:

- Reflect on my own position within the MML.
- Evaluate the changes made for enhancing access to the MML's digital collection.
- Deduce from the experience of a practical project new theoretical knowledge about Open GLAM in Social Movement Archives.
- Share with the MML control over the research process.

In October 2018, at the beginning of my PhD, I approached the MML whether they would be interested in a research collaboration that would co-investigate means of providing online access to the MML's collections. In the diagnosing phase I made myself familiar with the organizational culture, identified key players, and most importantly established trust. The diagnosing phase concluded with a focus group discussion about the MML's digitization objectives. We found consensus that it was the MML's priority to contribute with its digitized poster collection to the Social History Portal (SHP); a Europeana aggregator portal. In the planning stage I prepared the data

for the upload and designed a series of 6 evaluative online workshops for MML team members. The objectives of the workshops were that the participants:

- Reflect on the implications when collections are made available online through the SHP or Europeana.
- Develop criteria why to make certain collections available online (or why not) and set priorities.
- Learn about heritage copyright and its impact on the MML's digitization projects
- Understand how the SHP and Europeana are connected and their licensing conditions.

The action was completed with the successful poster upload to the SHP, and the workshops were conducted from September to October 2020. [1]

Contribution: Accounting on the limitations of the participatory approach and lessons-learned

Within heritage studies and DH, 'participation' has generally a positive connotation. However, the term's exact meaning remains often unclear (Flinn and Sexton, 2018: 626; Kidd, 2018: 201). Because the extent of participation is also not narrowly defined within the PAR methodology, it is necessary to assess critically what form of participation the research involved (Townsend, 2013: 101–03).

The participatory mode that took place in this project can be described as 'cooperative', where "local people work together with outsiders to determine priorities, responsibility remains with outsiders for directing the process" (Cornwall, 1996: 96). In this paper I am going to reflect on the factors that have shaped the mode of participation in this research. Specifically, I will address the following challenges and limitations, and how these could be mitigated in future DH projects:

- The possible mismatch between an academic research interest and the immediate priorities of a partner organization.
- The challenge to keep-up momentum and participants engaged due to academic administrative procedures.
- Unforeseeable circumstances, like the COVID-19 pandemic.
- The limitations of a three-year funded project for prolonged engagement and establishing mutual beneficial relationships (Herr and Anderson, 2015: 48–49; 150–57).

Bibliography

Cornwall, A. (1996). Towards participatory practice: participatory rural appraisal (PRA) and the participatory process. In Koning, K. de and Martin, M. (eds), *Participatory Research in Health: Issues and Experiences*. London: Zed Books ; NPPHCN, pp. 94–107.

Flinn, A. (2011). Archival Activism: Independent and Community-led Archives, Radical Public History and the Heritage Professions. *InterActions: UCLA Journal of Education and Information Studies*, **7** (2) https://escholarship.org/uc/item/9pt2490x (accessed 16 April 2021).

Flinn, A. and Sexton, A. (2018). Research on community heritage. *A Museum Studies Approach to Heritage*. 1st ed. Abingdon, Oxon ; New York, NY : Routledge, 2018. | Series: Leicester readers in museum studies: Routledge, pp. 625–39 doi:10.4324/9781315668505. https://www.taylorfrancis.com/books/9781315668505 (accessed 5 November 2021).

Herr, K. and Anderson, G. L. (2015). *The Action Research Dissertation: A Guide for Students and Faculty*. 2nd edition. Thousand Oaks: SAGE.

Hoyer, J. and Almeida, N. (2021). *The Social Movement Archive*. Sacramento, CA: Litwin Books.

Kemmis, S. and McTaggart, R. (2005). Participatory Action Research: Communicative Action and the Public Sphere. *The SAGE Handbook of Qualitative Research*. 3rd edition. Thousand Oaks: SAGE, pp. 559–603.

Kidd, J. (2018). Public Heritage and the Promise of the Digital. In Labrador, A. M. and Silberman, N. A. (eds), *The Oxford Handbook of Public Heritage Theory and Practice*. New York: Oxford University Press, pp. 198–208 doi:10.1093/oxfordhb/9780190676315.013.9.

Levenberg, L., Neilson, T. and Rheams, D. (eds). (2018). *Research Methods for the Digital Humanities*. New York, NY: Springer Berlin Heidelberg.

Ortolja-Baird, A. and Nyhan, J. (2021). Encoding the haunting of an object catalogue: on the potential of digital technologies to perpetuate or subvert the silence and bias of the early-modern archive. *Digital Scholarship in the Humanities* doi:10.1093/llc/fqab065.

Pickard, A. J. (2013). *Research Methods in Information*. 2nd edition. London: Facet.

Pringle, E. (2020). *Provisional Semantics: Addressing the Challenges of Representing Multiple Perspectives within an Evolving Digitised National Collection*. (Interim Report Foundation Projects Towards A National Collection) https://www.nationalcollection.org.uk/sites/default/files/2021-01/Provisional%20Semantics.pdf (accessed 11 October 2021).

Ruge, C., Wright, S. and Evans, J. (2016). Digital Dilemmas: a participatory investigation into developing

a digital strategy for a community archive. Melbourne http://www.vala.org.au/vala2016-proceedings/vala2016-session-13-ruge (accessed 28 November 2021).

Schuster, K. and Dunn, S. E. (eds). (2020). *Routledge International Handbook of Research Methods in Digital Humanities*. (Routledge International Handbooks). London New York: Routledge, Taylor & Francis Group.

Townsend, A. (2013). *Action Research: The Challenges of Understanding and Researching Practice*. Maidenhead, Berkshire; New York: Open University Press.

Wallace, A. (2020a). Words Mean Things (A Glossary). *Open GLAM* doi:10.21428/74d826b1.51566976. https://openglam.pubpub.org/pub/the-glossary (accessed 27 October 2021).

Wallace, A. (2020b). Introduction. *Critical Open GLAM: Towards [Appropriate] Open Access for Cultural Heritage* doi:10.21428/74d826b1.be9df175. https://openglam.pubpub.org/pub/introduction-to-critical-open-glam (accessed 29 October 2020).

Notes

1. I would like to thank everyone from the MML who participated in my research, as well as the MML's archivist and library manager Meirian Jump and my supervisors Professor Julianne Nyhan, Dr Antonis Bikakis and Dr Andrew Flinn for supporting me throughout the research process. The poster upload would not have been possible without Dr Donald Weber and the SHP team. Special thanks also to LaToyah Gill (Untamed Artists) and Matthew Lambert (British Library) for their guest workshop talks on copyright. Thank you to the anonymous reviewers of this conference paper for their feedback. This research was funded by a PhD studentship of the University College London.

Linked Data Approach for Studying Parliamentary Speeches and Networks of Politicians in Finland 1907-2021

Hyvönen, Eero
eero.hyvonen@aalto.fi
Aalto University, Finland; University of Helsinki (HELDIG), Finland

Leskinen, Petri
petri.leskinen@aalto.fi
Aalto University, Finland

Sinikallio, Laura
laura.sinikallio@helsinki.fi
University of Helsinki (HELDIG), Finland; Aalto University, Finland

Drobac, Senka
senkadrobac@aalto.fi
Aalto University, Finland; University of Helsinki (HELDIG), Finland

Tuominen, Jouni
jouni.tuominen@helsinki.fi
Aalto University, Finland; University of Helsinki (HELDIG), Finland

Elo, Kimmo
kimmo.elo@utu.fi
University of Turku

La Mela, Matti
matti.lamela@helsinki.fi
University of Helsinki (HELDIG), Finland

Koho, Mikko
mikko.koho@aalto.fi
Aalto University, Finland

Ikkala, Esko
esko.ikkala@aalto.fi
Aalto University, Finland

Tamper, Minna
minna.tamper@aalto.fi
Aalto University, Finland

Leal, Rafael
rafael.leal@aalto.fi
Aalto University, Finland; University of Helsinki (HELDIG), Finland

Kesäniemi, Joonas
joonas.kesaniemi@aalto.fi
Aalto University, Finland

Vision for a Paradigm Change

This paper presents the vision of publishing and using parliamentary data on the Semantic Web for Digital Humanities (DH) research, with a focus on studying parliamentary culture, language used, and networks of politicians (Hyvönen et al., 2021, 2022). First results of the Semantic Parliament project (ParliamentSampo, 2021) are presented:

1. Finnish parliamentary debates, totalling over 950 000 speeches and covering the whole history of the Parliament of Finland (PofF) 1907–2021, have been transformed into a 1) Speech Knowledge Graph (S-KG) and 2) into XML form using the new emerging Parla-CLARIN format (Parla-CLARIN, 2021) for international interoperability (Sinikallio et al., 2021).
2. A Prosopographical Knowledge Graph (P-KG) representing biographical data about the politicians during the same time span, using the event-based CIDOC CRM (CIDOC CRM, 2021) ontology, has been created and interlinked with the S-KG (Leskinen et al., 2021).
3. The datasets S-KG and P-KG were published as a Linked Open Data (LOD) service with a SPARQL endpoint and are used to study Finnish political culture and language, based on the speeches and networks of politicians (Hyvönen et al., 2021, 2022).
4. To demonstrate the usability of the new data infrastructure, a semantic portal *ParliamentSampo – Parliament of Finland on the Semantic Web,* targeted for researchers and the public is presented. The portal is based on the LOD service (3). The portal and LOD service will be opened for public use by the end of 2022.

Related Work

Parliamentary data are widely available for making political decision making transparent, and the data is used for linguistic and DH research. The paper first explains why publishing parliamentary data as LD makes sense, and discusses related projects in different countries (ParlaMint, 2021), including Canada, Italy, Latvia, Slovenia, UK, and the LinkedEP system (Van Aggelen et al., 2017) of the European Parliament. After this, the knowledge graphs (KG) of the ParliamentSampo system and their creation processes are presented and the benefits and challenges of the LOD approach are discussed, suggesting a paradigm shift in publishing and studying parliamentary data using semantic web technologies.

The Model and Implementation

Based on the Sampo-model (Hyvönen, 2021) and Sampo-UI framework (Ikkala et al., 2021), ParliamentSampo aggregates and enriches data from multiple data providers in addition to the PofF, and publishes the result in a LOD service, based on best practices of W3C (Heath & Bizer, 2011), including a SPARQL endpoint and additional LOD services, such as content negotiation. In addition, the 7-star LOD model (Hyvönen et al., 2014) extending the traditional 5-star model of Tim Berners-Lee with schema documentation and data validation is used. The LOD service can be used for direct DH analyses using its APIs and for creating ready-to-use applications for research. Data and application dissemination is supported using Docker containers.

Results and Evaluation

Feasibility of the ParliamentSampo approach is evaluated by showing how the SPARQL endpoint together with tools, such as YASGUI (Rietveld & Hoekstra, 2017) and Google Colab with Jupyter Notebooks, can be used for novel DH analyses and visualizations on parliamentary speeches and networks of politicians. This is the first time that all speeches of the PofF since it was established in 1907 are available as uniform data for DH research. We also introduce the new semantic portal "ParliamentSampo – Parliament of Finland on the Semantic Web" implemented on top of the LOD service. It is demonstrated how the portal can be used for analyzing political language in use in different times, their semantic content, and differences between prosopographical groups, such as female and male Members of the Parliament and different political parties. For this purpose, the speech texts have been enriched semantically with Named Entity Linking using FinBERT (2021), a Finnish language model based on Google BERT, by ontology-based keyword indexing using the automatic annotation tool Annif (Suominen, 2019), and by topic detection. Furthermore, network analyses of political networks using P-KG are presented using the Sparql2GraphServer tool (Leskinen et al., 2021). When using the portal, programming skills are not needed but data literacy. Finally, new possibilities and challenges of using linked data and ParliamentSampo in parliamentary studies are discussed and directions for further research are suggested.

The multidisciplinary work on ParliamentSampo has involved researchers in computer science, parliamentary studies, and linguistics at the University of Helsinki (HELDIG centre), Aalto University, and University of Turku, and is funded mostly by the Academy of Finland and EU project In/Tangible European Heritage. CSC – IT Center for Science, Finland, provided computational resources.

Bibliography

Van Aggelen, A., Hollink, L., Kemman, M., Kleppe, M., & Beunders, H. (2017). The Debates of the European Parliament as Linked Open Data. Semantic Web, 8(2), 271–281.

Parla-CLARIN. (Nov 24, 2021). Parla-CLARIN format: https://clarin-eric.github.io/parla-clarin/

CIDOC CRM. (Nov 24, 2021). CIDOC-CRM standard: https://cidoc-crm.org

FinBERT (Nov 24, 2021). Finnish BERT model: https://github.com/TurkuNLP/FinBERT

Heath, T., & Bizer, C. (2011). Linked Data: Evolving the Web into a Global Data Space.

Morgan & Claypool, Palo Alto, California.

Hyvönen, E., Tuominen, J., Alonen, M., & Mäkelä, E. (2014). Linked Data Finland: A 7-star Model and Platform for Publishing and Re-using Linked Datasets. In V. Presutti, E. Blomqvist, R. Troncy, H. Sack, I. Papadakis, & A. Tordai (Eds.), The Semantic Web: ESWC 2014 Satellite Events. ESWC 2014, pp. 226–230, Springer. https://link.springer.com/chapter/10.1007%2F978-3-319-11955-7_24

Hyvönen, E. (2021). Digital Humanities on the Semantic Web: Sampo Model and Portal Series. Submitted. https://seco.cs.aalto.fi/publications/2021/hyvonen-sampo-model-2021.pdf

Leskinen, P., Hyvönen, E., & Tuominen, J. (2021). Members of Parliament in Finland Knowledge Graph and its Linked Open Data Service. April. *Proceedings of SEMANTiCS – In the Era of Knowledge Graphs*, Amsterdam, Sept 6–9, 2021. https://seco.cs.aalto.fi/publications/2021/leskinen-et-al-mps-2021.pdf

Hyvönen, E., Sinikallio, L:, Leskinen, P., Drobac, S., Tuominen, J., Elo, K., La Mela, M., Koho, M., Ikkala, E., Tamper, M., Leal, R. & Kesäniemi, J. (2021). Semanttinen parlamentti: eduskunnan aineistojen linkitetyn avoimen datan palvelu ja sen käyttömahdollisuudet. Informaatiotutkimus, vol. 40, no. 2.

Hyvönen, E., Sinikallio, L:, Leskinen, P., Drobac, S., Tuominen, J., Elo, K., La Mela, M., Koho, M., Ikkala, E., Tamper, M., Leal, R. & Kesäniemi, J. (2022). Digital Parliamentary data in Action (DiPaDa 2022), Workshop at the 6th Digital Humanities in Nordic and Baltic Countries Conference, long paper, CEUR Workshop Proceedings, 2022. Forth-coming. https://seco.cs.aalto.fi/publications/2022/hyvonen-et-al-semparl-dhnb-2022.pdf

Ikkala, E., Hyvönen, E., Rantala, H., & Koho, M. (2022). Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces. Semantic Web – Interoperability, Usability, Applicability, 13(1), 69–84. https://doi.org/10.3233/SW-210428

Leskinen, P., Hyvönen, E. & Tuominen, J. 2021. Sparql2GraphServer: a Server-side Tool for Extracting Networks from Linked Data for Data Analysis. ISWC-Posters-Demos-Industry 2021 International Semantic Web Conference (ISWC) 2021: Posters, Demos, and Industry Tracks, CEUR Workshop Proceedings, Vol 2980. http://ceur-ws.org/Vol-2980/paper330.pdf

ParlaMint. (Nov 24, 2021). ParlaMint initiative homepage: https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora

Rietveld, L., & Hoekstra, R. (2017). The YASGUI family of SPARQL clients. *Semantic Web – Interoperability, Usability, Applicability, 8(3)*, 373–383. https://doi.org/10.3233/SW-150197

Sinikallio, L., Drobac, S., Tamper, M., Leal, R., Koho, M., Tuominen, J., La Mela, M., & Hyvönen, E. (2021). Plenary Debates of the Parliament of Finland as Linked Open Data and in Parla-CLARIN Markup. In: *3rd Conference on Language, Data and Knowledge (LDK 2021),* 1–17 . *OASICS, Schloss Dagstuhl, Leibniz-Zentrum fuer Informatik,* Germany. https://drops.dagstuhl.de/opus/volltexte/2021/14544/pdf/OASIcs-LDK-2021-8.pdf

ParliamentSampo. (Nov 24, 2021). Semantic Parliament project homepage: https://seco.cs.aalto.fi/projects/semparl/en/

Suominen, O. (2019). Annif: DIY automated subject indexing using multiple algorithms. Liber Quarterly, July. https://liberquarterly.eu/article/view/10732/11612

On the Road to Freedom: Network models of interviews of Czechoslovak respondents in the optics of audio-emotional analysis and computational linguistics.

Iashchenko, Anatoly Vladimirovich
yaschenko.anatoliy.ay@gmail.com
Universita, Roma, Itali

This Today, relations between Russia, on the one hand, and the Czech Republic and Slovakia, on the other, are rather problematic both politically and socio-culturally. Most of the disagreements arise against a backdrop of reinterpreted history, where any interpretation of domestic political events becomes an occasion for international scandals on both sides (I.M. Savelyeva, 2004; Hradilek, 2010).

The source for the study was an archive of video interviews from the Czech Institute for the Study of Totalitarian Regimes "Memory and History of Totalitarian Regimes".

In order to build network models and conduct comparative analysis between models derived by mixed methods of computational linguistics and emotional analysis

of audio recordings, 45 transcribed video interviews were selected from the source database.

The transcribed interviews were analysed in the MAXQDA 2020 PRO software. In the first step of the analysis of the interviews, codes were generated to break down the text. It is the cluster approach of breaking down the text into valences, meaning groups, that allows the linguistic analysis to be carried out so quickly and with such a high quality. From the first minutes of working with dissidents' memories, it becomes possible to study the field of historical consciousness.

Once the coding of the corpus of texts according to the results of the frequency analysis was completed, it became possible to create matrices of the frequencies of the coded keys in the text.

The next step in the interview process was to analyse the texts using the 'interactive word tree' method. This approach allowed us to identify "noise" and cleanse the matrix from unnecessary word combinations. Further analysis of the texts was carried out using content analysis tools: document portrait, document comparison table and code layout.

The resulting frequency tables and matrices for each interview were converted to CSV format and exported to the Gephi database, where network models of keywords in the form of graphs were constructed.

Speech cues are a natural way of human communication, involving both direct linguistic content (e.g., texts) and implicit paralinguistic information (e.g., speaker emotions). Although visual emotion identification is more advanced in modern science, emotional audio analysis is gaining popularity among researchers due to ever-improving algorithms and increasing accuracy.

The construction of network models of Czechoslovak dissidents' audio interviews based on emotional analysis allows not only a new perspective on computer analysis and visualisation methods as a researcher's tool, but also a wider disclosure of information potential of the sources under study. An important feature of this approach is the exclusion of the influence of emotional information of words.

A frequency classification table based on a two-dimensional vector model of emotion was created to conduct emotional analysis of audio interviews and subsequent construction of network models. Pyhton script was written and Librosa audio analysis package was used to investigate the acoustic features of the interviews. The audio files were analysed: fundamental frequencies (f0), spectral characteristics, chromaticity, amplitude, harmonic characteristics, MFCC, logarithmic spectra, etc. Also, to extract markers of vocal emotion from speech (Ma, Y., et al., 2019), logarithmic spectrum conversion to Mel scale followed by discrete cosine transformation was performed (Scherer, K.R., 2003).

The data were classified based on the audio recording spectra on a logarithmic scale for nine key emotions: sadness, fear, anger, frustration, excitement, disgust, happiness, surprise and neutral calm.

In the study, the first time the matrices of emotional models were constructed, an accuracy of 42% was achieved. In further analysis, difficulties were found in interpreting the emotions sadness and fear in female interviewees due to the lack of an extensive sound base of emotions in Czech. The algorithm was modified to incorporate expert judgement, and the accuracy of determination increased by more than 15%.

The tables of emotion analysis and emotion frequency were also translated into CSV files and uploaded to the Gephi software database to build network models in the form of graphs.

Results of comparative analysis of online models of audio-interviews and transcribed texts gave different representations of the same events, allowing the researcher, on the one hand, to expand the information potential of the source, and on the other hand, to interpret personal experiences and experiences of interviewees within a constructive group memory about the fall of the communist regime in Czechoslovakia.

The proposed methodology of analysis, based on the comparison of network models, allows not only a broader disclosure of the information potential of historical sources, but also provides a basis for modelling and studying the mechanisms of transmission of group memory.

Bibliography

I.M. Savelyeva, A.V. Poletaev. Social perceptions of the past: the types and mechanisms of formation. Preprint WP6/2004/07. - Moscow: State University of Higher School of Economics, 2004. 56 c.

Hradilek, Adam (ed.): Za vaši a naši svobodu. Torst, ÚSTR, Praha 2010.

Ma, Y.; Hao, Y.; Chen, M.; Chen, J.; Lu, P.; Košir, A. Audio-Visual Emotion Fusion (AVEF): A Deep Efficient Weighted Approach. Inf. Fusion 2019, 46, 184–192.

Scherer, K.R. Vocal Communication of Emotion: A Review of Research Paradigms. Speech Commun. 2003, 40, 227–256.

Modelling Gender Diversity – Research Data Representation Beyond the Binary

Illmer, Viktor J.
v.illmer@fu-berlin.de
Freie Universität Berlin, Germany

Poggel, Lisa
l.poggel@fu-berlin.de
Freie Universität Berlin, Germany

Diehr, Franziska
diehrf@rki.de
Robert Koch Institute, Germany

Drury, Lindsey
l.drury@fu-berlin.de
Freie Universität Berlin, Germany

Introduction

How may we model gender to account for its diversity while remaining simple enough to implement and query?

We address why gender diversity needs to be represented in databases, especially when confronted with historical sources. Analysing examples of gender modelling in established metadata schemata and descriptive data models, we propose a model that strikes a balance between an atomic and flexible ontology that returns valid results even for naïve data queries. We introduce the use of **gender qualifiers**, which allow nuanced statements on how the gender information was formulated. Use of the proposed modelling strategies are demonstrated following the Wikibase data model.[1]

Archiving the colonial concept of the gender binary

Historical research shows the gender binary to be a modern and colonial organising principle. Scholars have traced transgender identities throughout European history (Betancourt, 2020; Feinberg, 1996; Mauriello, 2019; Moyer, 2015) and examined how colonialism has affected Indigenous gender systems. Various communities in North America, Asia and Africa did not adhere to a gender binary before the imposition of colonial knowledge regimes (Amadiume, 2015; Hinchy, 2020; Neil & Garcia, 2009). Many Indigenous societies accepted genders beyond male and female (Cleves, 2014; Herdt, 1994; Mirandé, 2016; Roscoe, 1998; Slater & Yarbrough, 2011). Archives carry the imprint of colonial and heteronormative gender regimes (Arondekar, 2009; Ćosić et al., 2014). A critical modelling practice should therefore reflect the constructed nature of gender in the archival record and its limited temporal and geographical scope (see also Flanders & Jannidis, 2015, pp. 14–15).

Status quo: Representation of gender in established standards

Established metadata schemes do not adequately model the diversity of gender. Most standards use binary models, which only allow for the expressions **male** and **female**. The German National Library's Integrated Authority File, for example, includes the concepts **male**, **female** and **not known** (Deutsche Nationalbibliothek, 2019a, 2019b). In this way, it is similar to the system standardised by ISO 5218 (International Organization for Standardization, 2004). This vocabulary is inadequate even assuming it only catalogues purported biological sex, which are more diverse than this, e. g. intersexuality. Moreover, post-structural feminists argue that a distinction between sex and gender cannot be maintained (Butler, 2006).

Some binary models include an **'other'** category, such as the specification for vCard (Perreault, 2011). TEI guidelines include vCard and ISO 5218 vocabularies as examples to use for values of the element <sex> (Text Encoding Initiative Consortium, 2021). However, being identified with a category of **other** can be a stigmatising experience as well as one of othering (Kronk et al., 2021; Puckett et al., 2020). Othering also occurs in Wikidata's model, where a cis person's gender is recorded as **female**, while a trans person's gender is **transgender female**, thus drawing a distinction in gender between cis and trans individuals.

Introducing gender qualifiers

By distinguishing gender (with expressions such as **female**, **male**, **non-binary**, and more) from **gender modality**, a concept proposed by Ashley (pre-published) as an umbrella term that includes but is not limited to transgender and cisgender modalities, we make it possible for more precise information on a person's gender. Nonetheless, we avoid using the term gender **identity**, because historical sources seldom include gender self-identification, but instead are based on presumed or ascribed gender. In order to create a reliable data basis, we believe it indispensable to make explicit how the gender-specific information was formulated. We therefore propose extending gender information with qualifying statements. The terminology assembled in the GSSO ontology (Kronk & Dexheimer, 2020), which includes **assumed**,[2] **experienced** [3] (we propose using the term **self-identified** instead), **lived**,[4] and **recorded** [5] gender, appears especially suitable to this task. However, while the GSSO

lists non-Western genders such as hijra, muxe and X-gender as 'culturally specific' non-binary gender identities (Kronk & Dexheimer, 2020), we propose to instead place all gender instances on one level, countering the Eurocentrism in subsuming all genders under the concepts of non-binary and transgender (Kravitz, 2021).

Examples

The qualifier value **self-identified gender** should only be used if there is evidence of the gender identity given by the respective persons themselves. This is more often the case with contemporary, non-historical individuals (Figure 1). But historical exceptions exist: 18th-century preacher Public Universal Friend (Figure 2), for instance, who publicly identified as being reborn genderless and requested to be referred to without pronouns from 1776 on (Moyer, 2015, pp. 12–13).

Where there is no personal record available, the **self-identified gender** cannot be determined. Diné weaver and healer Hastiin Tł'a (1867–1937), for instance, was a nádleehi, a Diné Two-Spirit person. According to Naruszewicz (2016, pp. 2–3), weaving and the ceremonial duties of healers were gendered activities within the Diné tribe, and only nádleehi were allowed to follow both. The fact that Tł'a was both suggests that Tł'a's gender can be assumed as nádleehi (Figure 3). We suggest using **assumed gender** as a qualifier in this case and whenever there are only weak indicators (e.g. gendered pronouns, titles, occupations) available. This might also be the case when we are confronted with gender information derived from authority files, because there are no indicators available on how the information was determined in the first place.

Sometimes a gender may be recorded, but the reliability of the record is questionable or sources are contradictory. Throughout his life, 19th-century politician Murray Hall (Figure 4) presented male and registered a male name upon migrating to the US. When he died, his body was examined and his gender was posthumously legally determined to be female (Nelson, 2014, pp. 137–145). National media ridiculed him as a 'passing woman,' a term that carried an accusation of deception and trickery. In a case like this, we suggest users of the model settle on an approach that allows marking some statements as preferred over others. In the Wikibase context, so-called ranks may be used to emphasise or de-emphasise statements of the same property: The statement of recorded gender being female should be assigned a rank of **deprecated** to highlight that **lived gender** and **self-recorded gender** are more credible sources than the official record. Colonial records should be treated in the same fashion, especially when dealing with terms employed by colonial administrations, missionaries or ethnographers. Records from Christian missionaries, for instance, do not ascribe Tł'a a nádleehi gender but rather that of 'berdache,' a derogatory colonial term (Naruszewicz, 2016, p. 12; Figure 3).

Laxmi Narayan Tripathi		
instance of	Person	
gender	hijra	
	gender qualifier	self-identified gender
gender modality	transgender	
	described by source	Scroll.in (2016)

Yuu Watase		
instance of	Person	
gender	X-gender	
	gender qualifier	self-identified gender

Modelling examples for Laxmi Narayan Tripathi and Yuu Watase. Controversial hijra rights activist Tripathi has identified as transgender as well as hijra, which she has described as a kind of transgender identity. Due to its flexibility, our model allows entering both hijra and transgender as her gender modality. Manga artist Watase self-identifies being X-gender, but there is no information available on their gender modality.

Public Universal Friend		
instance of	Person	
gender	female	
	gender qualifier	recorded gender
	end date	1776
	agender	
	gender qualifier	self-identified gender
	start date	1776

Modelling example for Public Universal Friend. Before the announcement, the Friend's gender was recorded as female. However, the example is somehow exceptional: The Friend's announcement is not retroactive in scope (i. e. the Friend has not always identified as genderless), but is valid from 1776 onward. The flexibility of the model allows us to add date qualifiers to the statement and thereby accurately record this special case.

Hastiin Tł'a		
instance of	Person	
gender	nádleehi	
	gender qualifier	assumed gender
	berdache	
	gender qualifier	recorded gender
	rank	deprecated

Modelling example for Hastiin Tł'a. A qualifier specifies nádleehi as assumed gender and berdache as recorded gender. A Wikibase rank is used to mark the inaccurate and offensive nature of the colonial term. Whether it is appropriate to enter the term at all is subject to

discussion. With no information on how Tl'a self-identified, it is impossible to determine gender modality.

Murray Hall		
instance of	Person	
gender	male	
	gender qualifier	lived gender self-identified gender
	female	
	gender qualifier	recorded gender
	described by source	Nelson (2014)
	rank	deprecated

Modelling example for Murray Hall. Our entry for Murray Hall identifies both female and male as values of gender. A qualifier specifies female to be a recorded gender, taking into account the fact that Hall was recorded as female upon death. The male value carries two gender qualifier values: lived and self-identified gender. This gives credit to the fact that census and naturalisation records contain information provided by Hall in person. Given the hostile ridicule that his dead body was subjected to, it is impossible to determine whether Hall was, in fact, trans, leaving gender modality undefined.

Querying

One advantage of this approach is that a SPARQL query for persons and the value of their **gender** property returns valid and readable results. Using Wikibase, ranks may be applied to indicate the reliability of statements.[6] For instance, statements based on self-identification should be designated as preferred compared to statements that are based on records or assumptions. Queries would then, by default, only return the highest-ranked statements. Those wishing to go further in their analysis may explicitly query the gender qualifiers attached to a statement and receive the full range of nuances that the model is able to represent (Table 1).

Figure	Item	Gender	Gender qualifier	Rank	Gender modality
1	Laxmi Narayan Tripathi	hijra	self-identified gender	normal	transgender
1	Yuu Watase	X-gender	self-identified gender	normal	
2	Public Universal Friend	agender	self-identified gender	normal	
2	Public Universal Friend	female	recorded gender	normal	
3	Hastiin Tl'a	nádleehi	assumed gender	normal	
3	Hastiin Tl'a	berdache	recorded gender	deprecated	
4	Murray Hall	male	lived gender	normal	
4	Murray Hall	male	self-identified gender	normal	
4	Murray Hall	female	recorded gender	deprecated	

Table 1. SPARQL query results for all examples made thus far, including all gender statements and qualifiers. Sorted by order of appearance. Note the inclusion of deprecated gender statements, which would be withheld automatically when only querying for the gender property.

Closing remarks and prospects

In order to represent gender in the context of formalised data environments, we propose a modelling strategy that allows one to make explicit how the gender information was formulated in the first place. This enables researchers to make more informed choices. We hope that by distinguishing between different qualities of gendered information, our model can challenge essentialised notions and make cultural, geographical and temporal positions on gender explicit. It can thereby function as a tool that sparks discussion, uncertainty and scholarly self-reflection.

Bibliography

Amadiume, I. (2015). *Male Daughters, Female Husbands: Gender and Sex in an African Society*. 2nd ed. (Critique Influence Change). London: Zed Books.

Arondekar, A. (2009). *For the Record: On Sexuality and the Colonial Archive in India*. (Next Wave: New Directions in Women's Studies). Durham: Duke University Press, doi: 10.1515/9780822391029.

Ashley, F. 'Trans' is my gender modality: A modest terminological proposal. In Erickson-Schroth, L. (ed), *Trans Bodies, Trans Selves*. 2nd ed. Oxford University Press. [Preprint]. https://www.florenceashley.com/uploads/1/2/4/4/124439164/florence_ashley_trans_is_my_gender_modality.pdf.

Betancourt, R. (2020). *Byzantine Intersectionality: Sexuality, Gender, and Race in the Middle Ages*.

Princeton, NJ: Princeton University Press, doi: 10.1515/9780691210889.

Butler, J. (2006). *Gender Trouble: Feminism and the Subversion of Identity*. (Routledge Classics). New York: Routledge.

Cleves, R. H. (2014). Beyond the binaries in early America: Special issue introduction. *Early American Studies*, **12**(3). University of Pennsylvania Press: 459–68.

Ćosić, M., Dollinger, J., Isop, U. and Leibetseder, D. (2014). Gegenkulturelle Archive jenseits von Familie und Geschlecht. In Guggenheimer, J., Isop, U., Leibetseder, D. and Mertlitsch, K. (eds), *»When we were gender...« – Geschlechter erinnern und vergessen*, vol. 5. Bielefeld: transcript Verlag, pp. 245–72 doi: 10.1515/transcript.9783839423974.245.

Deutsche Nationalbibliothek (2019a). GND Gender https://d-nb.info/standards/vocab/gnd/gender_20191015.

Deutsche Nationalbibliothek (2019b). GND Ontology https://d-nb.info/standards/elementset/gnd_20191015.

Feinberg, L. (1996). *Transgender Warriors: Making History from Joan of Arc to RuPaul*. Boston, MA: Beacon Press.

Flanders, J. and Jannidis, F. (2015). Knowledge organization and data modeling in the humanities. https://nbn-resolving.org/urn:nbn:de:bvb:20-opus-111270.

Herdt, G. H. (ed). (1994). *Third Sex, Third Gender: Beyond Sexual Dimorphism in Culture and History*. New York: Zone Books.

Hinchy, J. (2020). *Governing Gender and Sexuality in Colonial India: The Hijra, c. 1850–1900*.

International Organization for Standardization (2004). *ISO/IEC 5218:2004*. https://www.iso.org/standard/36266.html.

Kravitz, M. (2021). Are pre-colonial genders inherently 'nonbinary' or 'transgender'? *An Injustice!* https://aninjusticemag.com/are-pre-colonial-genders-inherently-nonbinary-or-transgender-9667459e7574.

Kronk, C. A. and Dexheimer, J. W. (2020). Development of the Gender, Sex, and Sexual Orientation ontology: Evaluation and workflow. *Journal of the American Medical Informatics Association*, **27**(7): 1110–15 doi: 10.1093/jamia/ocaa061.

Kronk, C. A., Everhart, A. R., Ashley, F., Thompson, H. M., Schall, T. E., Goetz, T. G., Hiatt, L., et al. (2021). Transgender data collection in the electronic health record: Current concepts and issues. *Journal of the American Medical Informatics Association* doi: 10.1093/jamia/ocab136. https://academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocab136/6364772.

Mauriello, M. (2019). Corpi dissonanti: note su gender variance e sessualità. Il caso dei femminielli napoletani. *Archivio antropologico mediterraneo*. Dipartimento Culture e Società - Università di Palermo doi: 10.4000/aam.706.

Mirandé, A. (2016). Hombres mujeres: An indigenous third gender. *Men and Masculinities*, **19**(4). Los Angeles, CA: SAGE Publications: 384–409 doi: 10.1177/1097184X15602746.

Moyer, P. B. (2015). *The Public Universal Friend: Jemima Wilkinson and Religious Enthusiasm in Revolutionary America*. Ithaca, N.Y: Cornell University Press, doi: 10.7591/9781501701450.

Naruszewicz, C. J. (2016). Beyond binary: Navajo alternative genders throughout history. https://hdl.handle.net/11244/325130.

Neil, C. and Garcia, J. (2009). *Philippine Gay Culture: Binabae to Bakla, Silahis to MSM*. Hong Kong: Hong Kong University Press.

Nelson, L. (2014). Reanimating archiving/archival corporealities: Deploying 'big ears' in de rigueur mortis intervention. *QED: A Journal in GLBTQ Worldmaking*, **1**(2). Michigan State University Press: 132–59 doi: 10.14321/qed.1.2.0132.

Perreault, S. (2011). *VCard Format Specification*. (Request for Comments). RFC Editor doi: 10.17487/RFC6350. https://rfc-editor.org/rfc/rfc6350.html.

Puckett, J. A., Brown, N. C., Dunn, T., Mustanski, B. and Newcomb, M. E. (2020). Perspectives from transgender and gender diverse people on how to ask about gender. *LGBT Health*, **7**(6): 305–11 doi: 10.1089/lgbt.2019.0295.

Roscoe, W. (1998). *Changing Ones: Third and Fourth Genders in Native North America*. 1. ed. New York: StMartin's Press.

Scroll.in (2016). Why I chose to become a hijra: Laxmi in her own words Text *Scroll.In* http://scroll.in/article/814182/why-i-chose-to-become-a-hijra-laxmi-in-her-own-words.

Slater, S. and Yarbrough, F. A. (eds). (2011). *Gender and Sexuality in Indigenous North America, 1400–1850*. Columbia, S.C: University of South Carolina Press.

Text Encoding Initiative Consortium (2021). TEI element <sex> *P5: Guidelines for Electronic Text Encoding and Interchange* https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-sex.html.

Notes

1. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy in the context of the Cluster of Excellence Temporal Communities: Doing Literature in a Global Perspective – EXC 2020 – Project ID 390608380.
2. assumed gender: 'The assumption of a person's gender based on any predefined characteristics and biases.'

3. experienced gender: ‘The gender with which a person identifies.’
4. lived gender: ‘The gender in which an individual lives their everyday life. This is usually a combination of performance of expected gender roles and gender expression, and may not represent a person’s actual gender.’
5. recorded gender: ‘Gender as recorded by any of various institutions, organisations, writings, etc.’
6. Although we chose Wikibase as a suitable environment for this task, we wish to highlight that our approach may also be adapted to other systems.

Abstractness/ Concreteness as Stylistic Features for Authorship Attribution

Ivanov, Lubomir
livanov@iona.edu
Iona College, United States of America

Introduction

We present early results from a broad project investigating the usefulness of abstractness/concreteness as stylistic features for authorship attribution. The concreteness of a word/phrase refers to our sensory ability to perceive or experience the object/phenomenon described by that word/phrase. Abstractness is the opposite of concreteness. Abstractness and concreteness have been studied from a philosophical, psychological, neuro-physiological, linguistic, and literary perspectives. A significant amount of work has been performed on the computational aspects of concreteness/abstractness recognition and classification, including metaphor-, hyperbole- and idiom detection, rating, and processing (Spreen and Schulz, 1966; Paivio et al, 1968; Coltheart, 1981, Birke and Sarkar, 2006 & 2007; Li and Sporleder, 2009; Sporleder and Li, 2009; Turney et al, 2011; Kwong, 2011, Assaf et al, 2013; Neuman et al, 2013; Shutova et al, 2013; Brysbaert et al, 2014; Tsvetkov et al., 2014; Klebanov et al, 2015; Rei et al, 2017; Wu et al, 2018, Gao et al, 2018; Reif et al, 2019). There have been attempts to quantify the notions of concreteness and abstractness (Spreen and Schulz, 1966; Paivio et al, 1968, Brysbaert et al, 2014). In (Brysbaert et al, 2014), a collection of 37058 words and 2896 two-word phrases was rated for concreteness by 4000 evaluators (approximately 25 evaluators per word). The rating uses a 5-point scale, where lower values indicate abstractness and higher values - concreteness. For each word/phrase, the mean rating and the standard deviation (i.e. evaluator agreement) were recorded. The Brysbaert dictionary has become a standard in many studies and has been used to train machine learning models to automatically rate word/phrase abstractness (Köper and Schulte Im Walde, 2016, Maudslay et al, 2020).

We are interested in the usefulness of abstractness/ concreteness as stylistic features for authorship attribution. We present a new attribution methodology based on the Brysbaert dictionary. We employed the methodology to perform attribution experiments using the Reuters-RCV1 dataset (Lewis et all, 2014, NIST). The results show promise and warrant further studies.

Attribution Methodology

Our strategy is to distribute the words (and word-pairs) from each document into categories based on the part of speech (PoS) roles of the words and the agreement of the evaluators as to the word’s abstractness. PoS information is important because words carry different levels of abstractness based on the context. For example, “toy” is concrete as a noun (mean 4.93, standard deviation 0.38 in the Brysbaert dictionary) but less concrete as a verb (mean 2.3, standard deviation 1.17). Similarly, “chestnut”, as a noun, is very concrete, but as a color adjective – quite abstract. We use the following PoS categories: jj, jjr, jjs, nn, nnp, nns, rb, rbr, rbs, vb, vbd, vbg, vbn, vbp, vbz. These are based on the standard Penn Treebank tags (Santorini, 1990, Penn Treebank Tags). We also have a category “wp” for word pairs.

The standard deviation (SD) of the evaluators’ ratings in the Brysbaert dictionary is another indicator of abstractness. A scan through the dictionary reveals that SD is smaller for nouns (average 1.10), larger for adjectives (average 1.218), and even larger for verbs (average 1.253). In general, SD is larger for abstract words. We define four SD classes: “very narrow” (SD<0.5), “narrow” (0.5≤SD<1.0), “wide” (1.0≤SD<1.5) and “very wide” (SD≥1.5).

Finally, we combine the PoS and SD classes above to obtain 64 abstractness classes (e.g. “jj_narrow”, “vb_wide”, etc.). Our algorithm constructs a 64-dimensional vector for each text in the corpus by mapping each word in the text to one of the 64 classes above and averaging the mean abstractness ratings of all words mapped to the same category. We exclude the most (universally) common 25 nouns, 50 verbs and their tenses, 50 adjectives, and 35 adverbs to avoid skewing the results by frequently used words. The generated vectors are stored as WEKA (Hall et al, 2009) files and used for training WEKA classifiers.

Experiments and Results

Our experiments were based on randomly selected 20-, 15-, and 10-author subsets of the Reuters-RCV1 corpus. We PoS-tagged all words in each text in the corpus using the Stanford PoS tagger. Next, we generated an abstractness vector for each file and trained three WEKA classifiers - a support vector machine with sequential minimal optimization (SMO), a multilayer perceptron (MP), and a random forest classifier (RF). We used leave-one-out cross-validation. Table 1 shows the averages of the results from all three classifiers for abstractness as well as for a set of traditional stylistic features.

Table 1: Average-of-3-classifiers accuracies for different features

	20 Authors	15 Authors	10 Authors
Character-2-Grams	69.8%	75.2%	83.6%
Character-3-Grams	59.5%	64.6%	77.4%
M-W function words	45.3%	56.0%	64.7%
First-word-in-sentence	38.0%	47.4%	63.5%
Coarse POS Tagger	54.3%	65.5%	72.8%
Prepositions	38.1%	44.5%	55.3%
Suffices	64.1%	73.1%	80.6%
Vowel-initiated words	49.4%	60.0%	69.4%
Word-2-grams	43.5%	54.4%	66.9%
Abstractness	60.10%	71.5%	79.0%

Table 2: Abstractness confusion matrix for 10 authors

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j   <-- classified as
 37  1  0  2  1  4  1  3  1  0 | a = MarcelMichelson
  4 33  3  1  2  0  4  2  0  1 | b = DarrenSchuettler
  2  2 44  0  0  0  1  1  0  0 | c = LydiaZajc
  1  0  0 42  1  1  0  0  5  0 | d = DavidLawder
  1  3  0  0 38  0  1  0  0  7 | e = GrahamEarnshaw
  2  0  0  2  0 41  0  2  2  1 | f = RogerFillion
  0  1  0  0  1  0 48  0  0  0 | g = FumikoFujisaki
  0  2  0  1  1  1  3 33  0  1 | h = BernardHickey
  5  0  0  6  0  0  1  2 35  1 | i = ToddNissen
  0  0  0  0  3  2  1  0  0 44 | j = PeterHumphrey
```

While not the top performing stylistic feature, abstractness outperforms most standard features including function words. The confusion matrix in Table 2 – typical across all abstractness experiments - demonstrates that precision, recall, and F-measure are high for most authors. Some authors exhibit particularly high precision and recall (e.g. Lydia Zajc: 0.936/0.880, Fumiko Fujisaki: 0.800/0.960, etc.). This indicates that some authors use abstraction/concreteness in unique ways, but more work is needed to determine the patterns of abstractness that set these authors apart.

Conclusion and Future Work

We presented an authorship attribution methodology based on the use of abstractness/concreteness ratings. The early results show promise but further research is needed:

- Perform further experiments with different datasets: We have several datasets available – a Victorian authors dataset, a poetry dataset based on Project Guttenberg, an 18 th century American/British documents dataset, and a 19 th century American literary dataset.
- Explore automatic methods for rating abstractness/concreteness.
- Investigate the issue of abstractness/concreteness complementarity and its impact on authorship attribution.
- Consider the abstractness of longer multiword phrases in authorship attribution.
- Explore the use of metaphors in attribution.
- Study abstractness patterns in literary works and their impact on the attribution accuracy.

Bibliography

Assaf D., Neuman Y., Cohen Y., Argamon S., Howard N., Last M., Frieder O., Koppel M. (2013). Why "dark thoughts" aren't really dark: A novel algorithm for metaphor identification. In *Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB), 2013 IEEE Symposium on*, pages 60–65. IEEE

Brysbaert, M., Warriner, A.B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46, 904-911.

Birke J., Sarkar A. (2006). A Clustering Approach for the Nearly Unsupervised Recognition of Nonliteral Language. In *Proceedings of the 11th Conference of the European Chapter of the ACL*, pages 329–336, Trento, Italy.

Birke J., Sarkar A. (2007). Active Learning for the Identification of Nonliteral Language. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 21–28, Rochester, NY.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology,* 33, 497-505.

Gao G., Choi E., Choi Y., Zettlemoyer L. (2018). Neural metaphor detection in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 607–613, Brussels, Belgium. Association for Computational Linguistics.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I., (2009). *The WEKA Data Mining Software: An Update; SIGKDD Explorations*, Volume 11, Issue 1

Klebanov B., Leong C., Flor M. (2015). Supervised word-level metaphor detection: Experiments with concreteness and reweighting of examples. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 11–20, Denver, CO. ACL.

Köper, M & Schulte Im Walde, S. (2016). Automatically Generated Affective Norms of Abstractness, Arousal, Imageability and Valence for 350 000 German Lemmas. *LREC'16*.

Kwong O.Y. (2011), Measuring Concept Concreteness from the Lexicographic Perspective, *25th Pacific Asia Conference on Language, Information and Computation, pages 60–69*

Lewis, D. D.; Yang, Y.; Rose, T.; and Li, F. (2004). RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research*, 5:361-397

Maudslay R., Pimentel T., Cotterell R., Teufel S. (2020). Metaphor Detection using Context and Concreteness, *Proceedings of the Second Workshop on Figurative Language Processing*, pages 221—226.

Neuman Y., Assaf D, Cohen Y., Last M., Argamon S., Howard N., Frieder O. (2013). Metaphor identification in large texts corpora. *PloS One*, 8(4):e62343.

NIST: https://trec.nist.gov/data/reuters/reuters.html

Paivio, A., Yuille. J., Madigan, S. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of experimental psychology*, 76(1p2):1

Penn Treebank Tags: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Rei M., Bulat L., Kiela D., Shutova E. (2017). Grasping the finer point: A supervised similarity network for metaphor detection. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1537–1546.

Santorini B. (1990) "Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision)"

Shutova E., Teufel S., Korhonen A. (2013). Statistical Metaphor Processing. *Computational Linguistics*, 39(2):301–353.

Sporleder C., Li L. (2009). Unsupervised Recognition of Literal and Non-Literal Use of Idiomatic Expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 754–762, Athens, Greece.

Spreen O., Schulz. R., (1966). Parameters of abstraction, meaningfulness, and pronunciability for 329 nouns. *Journal of Verbal Learning and Verbal Behavior*, 5(5):459–468.

Tsvetkov Y., Boytsov L., Gershman A., Nyberg E., Dyer C. (2014). Metaphor Detection with Cross-Lingual Model Transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 248–258.

Tsvetkov Y., Mukomel E., Gershman A. (2013). Cross-lingual metaphor detection using common semantic features. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 45– 51, Atlanta, Georgia. ACL.

Turney P., Neuman Y., Assaf D., Cohen Y. (2011). Literal and Metaphorical Sense Identification through Concrete and Abstract Context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, UK.

Wu C., Wu F., Chen Y., Wu S., Yuan Z., Huang Y. (2018). Neural metaphor detecting with CNN-LSTM model. In *Proceedings of the Workshop on Figurative Language Processing*, pages 110–114, New Orleans, Louisiana. ACL.

Studying Signs of Use in Medieval Manuscripts: data collection through annotations

Johnson, David F.
djohnson@fsu.edu
Florida State University

Girard, Paul
paul@ouestware.com
OuestWare, Nantes, France

Simard, Benoît
benoit@ouestware.com
OuestWare, Nantes, France

Written signs of use

How do we know that a medieval manuscript has actually been read? Early medieval readers of manuscripts rarely "signed" their books or included an ex libris. The modern researcher interested in whether a given manuscript was actually read looks for explicit signs of use left behind by readers: traces of interactions with the contents of the books they were reading, more often than not pen in hand. Such traces comprise a wide range of interventions.

Despite his anonymity, perhaps the most famous reader of Old English texts is the so-called "Tremulous Hand of Worcester," a monk of the Cathedral Priory at Worcester, who left his unmistakable mark on a significant number of Worcester manuscripts containing Old English in the first half of the thirteenth century. These manuscripts, however, were written in a form of English not readily comprehensible to him. He set out to learn to read it, and in the process left a wide range of signs of use in the books he used (some twenty manuscripts in total). The Tremulous Hand used a system of punctuation interventions in his work—distinct from the original punctuation applied by the Old English scribes who wrote the manuscripts—as a tool for construing that much older form of English (Harlow, 1959; Parkes, 1993; Johnson, 2006; Reimer, 2015; Johnson, 2021a). It was a kind of 'interpretative pointing', whereby he sounded out the Old English and used these punctuation interventions to work out the syntax and meaning of the text he was learning to read. Such interventions are also among the clearest indications of the parts of these texts that he was truly interested in; tracing them and subsequently mapping

them out over the texts in which they appear reveals patterns of interest and is, we argue, like looking over the shoulder of the 13th-century scribe as he read these ancient books.

Annotations as a collection tool

In order to track, compile, and facilitate quantitative analysis of thousands of signs of use, we need a tool capable of capturing and indexing those written reading interventions from digital versions of the manuscripts. Annotations can be used to capture signs of use by delimiting and adding metadata to the part of the image where the sign appears. Although this may resemble the broader field of transcription, it differs as only a very small part of the text (the reading signs) are in scope; the aim is not to transcribe from pixels to text but to describe in an analytical perspective.

Such a method proposes to use annotations as a collection tool. By adding specific metadata to the annotated reading signs captured, annotating the studied manuscript creates a research dataset to be used in quantitative analysis.

The annotations are here a perfect device to bridge the close reading in context (the captured set of pixels) with the "distant reading" analysis done later on the metadata collected across the corpus of annotations (Moretti, 2013).

Existing manuscript annotation tools focus on different tasks, such as transcription and curation, or are closely linked to one corpus (Nagasaki et al., 2016; Roddis and Cogapp, 2018; Almas et al., 2018). Our ambition is to fulfil our methodological need with a generic IIIF image annotation tool dedicated to collecting data for "distant reading" analysis.

Tremulator application

All of Johnson's work on the Tremulous Hand's interventions published thus far were facilitated by the first iteration of the Tremulator application (named after the scribe mentioned above) (Johnson, 2019; Johnson, 2021b). Development of Tremulator 2.0 was made possible by a grant from Florida State University. Tremulator 2.0 has been designed to fit our research project while at the same time targeting a wider range of applications. We set out to combine existing open source technologies related to image annotations to create a generic service for data collection.

We chose to use the IIIF standards to facilitate accessibility to high resolution pictures and to benefit from existing collections already using this format around the world by employing an extended version of the existing IIIF-leaflet plugin.

For all the benefits of IIIF compatibility, it is not unusual to have to work with images procured by the research team via other means. Thus the Tremulator contains a cantaloupe server to serve local files under IIIF format. By using IIIF as the common ground for image retrieval, Tremulator's collections can mix different points of origins being a distant IIIF server or local files.

The annotation shapes creation is handled as geojson geometries. To enable complex annotation description schemas, we developed a JSON-schema creation form. This form creates annotation schemas by allowing the combining of numeric, textual and list of values fields. To allow reviewing the schemas on the go, new fields can be created at any time of data collection.

The collected data can be explored in a dedicated analysis page or exported as CSV to drive further analysis in other software.

The ability to create multiple custom data schemas for annotations, to use the IIIF format for both local or remote images and to browse annotation picture snippets are Tremulator's key features compared to other similar tools such as Recogito [1].

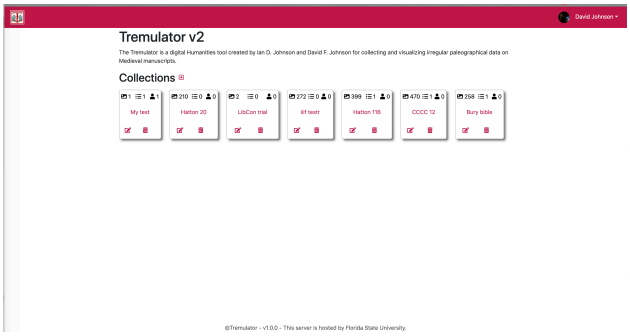
Discussion

We hope to have contributed to the broad community of image analysis with software that facilitates rich data collection on images, by targeting a more generic application than the one our research project actually needs.

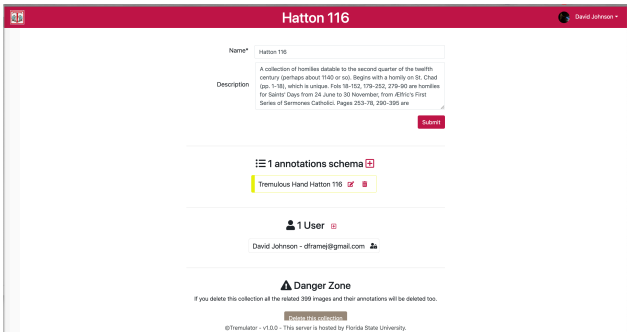
So far we have focused on creating a usable web application for data collection and export. Future development will target more advanced visual analytics, such as heatmaps of image pages showing the positions of annotations.

This tool will be of particular interest to scholars working in the fields of paleography, codicology, and other aspects of manuscript studies, but it should also appeal more broadly to anyone with an interest in the digital humanities, book history, art history, semiology, archeology and architecture, and any number of other fields for which the collection and visualization of irregular data is desirable.

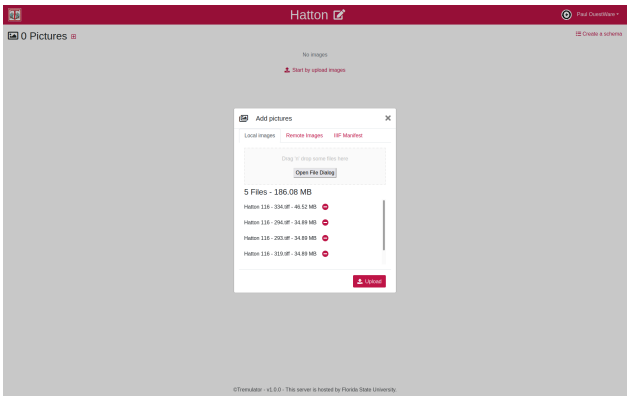
Figures



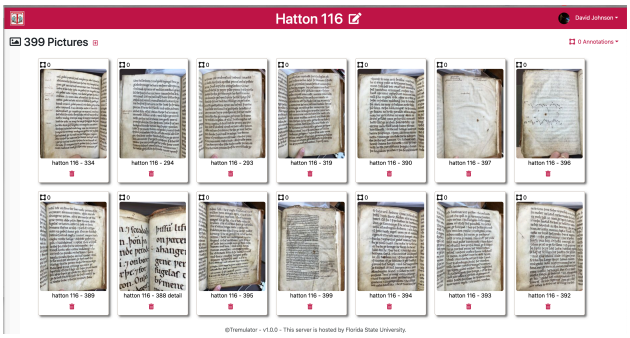
Tremulator's Home page showing the list of one users' collections



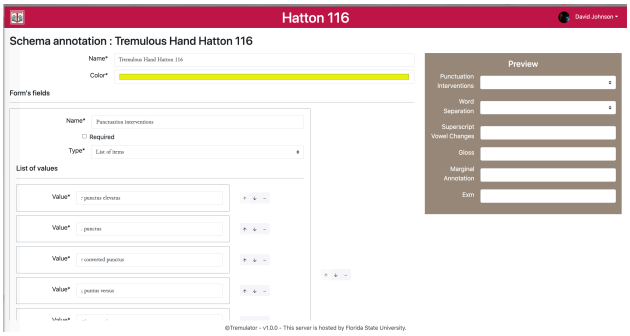
Collection administration page



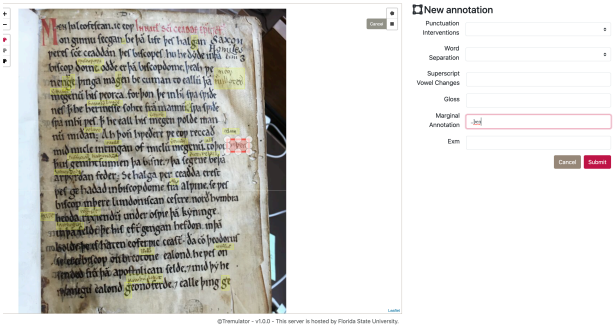
Add pictures to a collection by uploading local images, providing remote images URLs and/or importing a IIIF Manifest



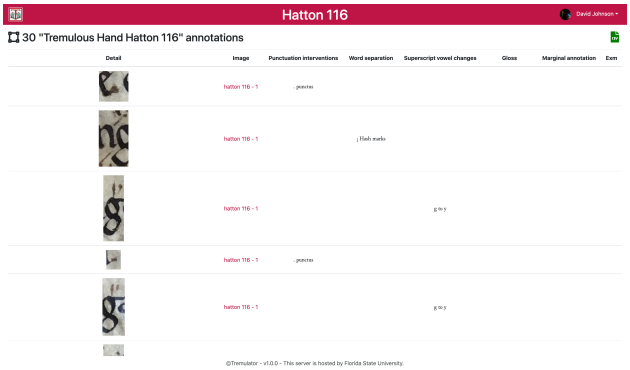
Collection pictures gallery



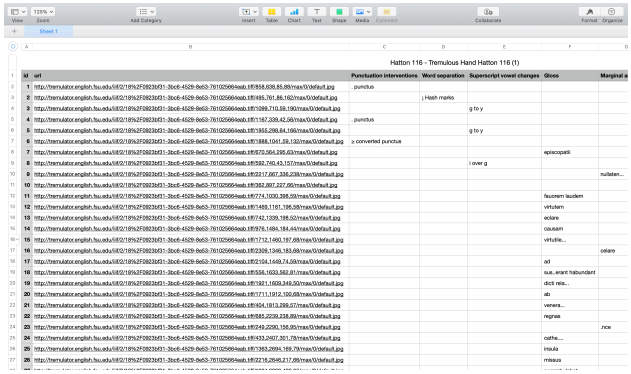
Annotations' schema edition



Annotation edition



Annotations data table withsnippets and CSV export



Hatton 116 - Tremulous Hand Hatton 116 (1)

id	url	Punctuation interventions	Word separation	Superscript vowel changes	Gloss	Marginal a
1	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/858,838,85,85/max/0/default.jpg	, punctus				
2	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/495,761,86,162/max/0/default.jpg		; Hash marks			
3	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/1089,710,69,190/max/0/default.jpg			g to y		
4	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/1167,339,42,56/max/0/default.jpg	, punctus				
5	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/1955,298,84,190/max/0/default.jpg			g to y		
6	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/1884,1061,69,130/max/0/default.jpg	; converted punctus				
7	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/870,964,295,63/max/0/default.jpg				episcopatli	
8	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/592,740,43,157/max/0/default.jpg			i over g		
9	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/2217,997,336,238/max/0/default.jpg					nullatenus
10	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/382,897,227,65/max/0/default.jpg					
11	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/774,1030,398,59/max/0/default.jpg				faucrem laudem	
12	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/1468,1161,196,58/max/0/default.jpg				virtutem	
13	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/742,1339,198,59/max/0/default.jpg				eclare	
14	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/976,1484,184,44/max/0/default.jpg				causam	
15	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/712,1490,197,69/max/0/default.jpg				virtutis...	
16	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/2309,1346,183,68/max/0/default.jpg					celare
17	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/2104,1449,74,59/max/0/default.jpg				od	
18	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/556,1633,582,81/max/0/default.jpg				sue...erant habundant	
19	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/1821,1609,349,50/max/0/default.jpg				dicit rela...	
20	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/1711,1912,100,69/max/0/default.jpg				ab	
21	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/404,1813,299,57/max/0/default.jpg				venera...	
22	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/685,2230,238,89/max/0/default.jpg				regnas	
23	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/249,2290,156,95/max/0/default.jpg					nce
24	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/433,2497,301,79/max/0/default.jpg				cathe...	
25	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/1363,2694,169,79/max/0/default.jpg				insula	
26	http://tremulator.english.fsu.edu/iiif/2/18%2F0803d51-3bc6-4529-8e53-76102f566eab.tif/2216,2646,217,66/max/0/default.jpg				miseus	

Annotations data as CSV in a spreadsheet software

Bibliography

Almas, B., Khazraee, E., Miller, M. T. and Westgard, J. (2018). Manuscript Study in Digital Spaces: The State of the Field and New Ways Forward. *Digital Humanities Quarterly*, **012**(2).

Harlow, C. G. (1959). Punctuation in Some Manuscripts of Ælfric. *Review of English Studies*, **10**: 1–19.

Johnson, D. F. (2006). Who Read Gregory's Dialogues in Old English? In Wilcox, J. and Magennis, H. (eds), *The Power of Words: Anglo-Saxon Studies Presented to Donald G. Scragg on His Seventieth Birthday*. Morgantown: West Virginia University Press, pp. 173–204.

Johnson, D. F. (2019). The Micro-Texts of the Tremulous Hand of Worcester: Genesis of a Vernacular liber exemplorum. In Lenker, U. and Kornexl, L. (eds), *Anglo-Saxon Micro-Texts*. Berlin, Boston: De Gruyter, pp. 225–66.

Johnson, D. F. (2021a). MS CUL Kk 3.18 and the Tremulous Hand of Worcester. In Brady, L. (ed), *Essays on Old English Literature in Honor of J.R. Hall*. MRTS, University of Arizona Press.

Johnson, D. F. (2021b). The Transmission and Reception of Alfredian 'Apocrypha'. In Breay, C. and Story, J. (eds), *Manuscripts in the Anglo-Saxon Kingdoms: Cultures and Connections.* Dublin: Four Courts Press, pp. 98–107.

Moretti, F. (2013). *Distant Reading*. London: Verso.

Nagasaki, K., Tsuda, T., Muller, C. and Shimoda, M. (2016). Tagging on Buddhist Images via IIIF and TEI encoding. *TEI Conference and Members' Meeting*. pp. 141–43.

Parkes, M. B. (1993). *Pause and Effect: An Introduction to the History of Punctuation in the West*. Berkeley: University of California Press.

Reimer, S. (2015). Manuscript Studies: Paleography: Punctuation http://www.ualberta.ca/~sreimer/ms-course/course/punc.htm (accessed 27 August 2015).

Roddis, T. and Cogapp, U. (2018). Making metadata into meaning: digital storytelling With IIIF. *By N. Proctor & R. Cherry. MW18: MW*.

Notes

1. Recogito, an initiative of Pelagios Commons, http://recogito.pelagios.org/ (accessed 11 April 2022)

Standards, the Standards-Making Process, and their Relevance to Stylometry

Juola, Patrick

juola@mathcs.duq.edu
Duquesne University, United States of America

The stakes have never been higher for stylometry applications and research. In addition to investigating academic literary and historical questions (Burrows, 2003), stylometrists are increasingly called upon to provide forensic evidence. Stylometry is being employed to resolve legal issues such as murder (Chaski, 2007; Grant, 2012), fraud (McMenamin, 2011), and asylum applications (Juola, 2014). Judges and juries need straightforward and understandable answers to questions such as "did the defendant write this email?" "is this will forged?" or "is this suicide note genuine?" (Chaski, 2007; Ainsworth & Juola, 2018) With questions of justice, substantial financial outcomes, and individual safety hanging in the balance, the need for accuracy in stylistic analysis is crucial.

Accuracy in forensic science has been recognized as a "crisis" (National Research Council, 2009) in that much of it relies on "questionable or questioned science" (National Research Council, 2009) with little empirical support. For example, forensic odontology (dentistry) simply doesn't work. (PCAST, p. 3; Pilkington, 2022) The (US) President's Council of Advisors on Science and Technology (PCAST) discussed both "foundational validity" and "validity as applied" as absolute requirements for forensic evidence, and further focused on the need for standards of practice to evaluate whether these requirements have been met. The American Academy of Forensic Sciences (AAFS) Standards Board was established in 2015 to provide ``high quality science-based consensus forensic standards'' in a variety of disciplines, including forensic document examination. While the (US) government does not typically set standards

or mandate their use, cleaving to standards can enhance reliability, credibility, and transparency of forensic evidence.

As peer reviewers, the stylometric community is used to evaluating the validity of individual stylometric analyses on a paper-by-paper basis. We are familiar with questionable practices that may produce inaccurate or untrustworthy results, and can recommend changes for better outcomes. For instance, machine learning methods can easily "overfit" the training data at the expense of accuracy on the actual data of interest, hence the need for validation on representative data prior to analysis. At the same time, we recognize that scholarly disciplines are continually changing and the best practices of twenty years ago, while still good practices, may have been overtaken by new and improved practices. For example, improved classification methods such as deep learning may outperform simple feature comparison methods such as Burrow's Delta (2003), but at the same time may require an impractical amount of training data for real-world problems, and may also be too confusing to explain to a judge and a jury.

However familiar these points are to DH practitioners, legal experts cannot be expected to know or understand them. Formal standards and best practices can provide guidance to the general public in recognizing and excluding clearly unacceptable work.

This paper discusses the standards-making process, including the language of standards, the creation and publication process, and the role of standards in interpreting forensic evidence, in order to promote discussion of accountability and accuracy in high-stakes application of stylometry. We specifically highlight the work of Rudman (2005; 2012) and Juola (2015) on stylometric accuracy and the handling of documents to maximize accuracy. We address the nature of unreliable analyses and appropriate methods to exclude less dependable techniques in situations with profound legal, financial, and human rights implications, while still allowing for scholarly research and exploration. The consequences of bad stylometry are significant; it would be valuable to draw up a list of guard rails and red lines that can mark an analysis as untrustworthy and therefore not to be accepted or relied upon.

Bibliography

Burrows, John. "Questions of authorship: attribution and beyond: a lecture delivered on the occasion of the Roberto Busa Award ACH-ALLC 2001, New York." *Computers and the Humanities* 37, no. 1 (2003): 5-32.

Chaski, Carole. "The keyboard dilemma and authorship identification." In *IFIP International Conference on Digital Forensics*, pp. 133-146. Springer, New York, NY, 2007.

Grant, Tim. "TXT 4N6: method, consistency, and distinctiveness in the analysis of SMS text messages." *Journal of Law & Policy* 21 (2012): 467-494.

McMenamin, Gerald. "Declaration of Gerald McMenamin."*Ceglia v. Zuckerberg and Facebook, WD 2012 WL 1392965* (W.D.N.Y). (2012). Available online athttp://www.scribd.com/doc/67951469/Expert-Report-Gerald-McMenamin.

Juola, Patrick., Stylometry and Immigration: A Case Study, *Journal of Law & Policy* 21 (2012): 287-298

Ainsworth, Janet, and Patrick Juola. "Who wrote this: Modern forensic authorship analysis as a model for valid forensic science." *Washington University Law Review* 96 (2018): 1159.

National Research Council. *Strengthening Forensic Science in the United States: A Path Forward*. Washington, DC:National Academies Press, 2009.

President's Council of Advisors on Science and Technology.*REPORT TO THE PRESIDENT Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*, Washington DC, 2016.

Pilkington, Ed. "A bite mark, a forensic dentist, a murder: How junk science ruins innocent lives." *The Guardian, 28 April 2022*. Available online athttps://www.theguardian.com/us-news/2022/apr/28/forensics-bite-mark-junk-science-charles-mccrory-chris-fabricant

Rudman, Joseph. "Unediting, de-editing, and editing in nontraditional authorship attribution studies: With an emphasis on the canon of Daniel Defoe." *The Papers of the Bibliographical Society of America* 99, no. 1 (2005): 5-36.

Rudman, Joseph. "The State of Non-Traditional Authorship Attribution Studies—2012: Some Problems and Solutions." *English Studies* 93, no. 3 (2012): 259-274.

Juola, Patrick. "The Rowling case: A proposed standard analytic protocol for authorship questions." *Digital Scholarship in the Humanities* 30, no. suppl_1 (2015): i100-i113.

Quantitative Analysis of Changes in Race and Role of Characters in Hollywood Movies over Time

Kawase, Akihiro
kawase@dh.doshisha.ac.jp
Doshisha University, Japan

Isogai, Kana
isogai.kana@dh.doshisha.ac.jp
Doshisha University, Japan

Introduction

Movies are a medium that influences viewers' beliefs and values. Behm-Morawitz and Mastro (2008) demonstrated that the image of "mean girls" in many movies created a negative stereotype of female friendships. In addition, males who watched more movies targeting teenagers, tended to display more negative attitudes toward female friendships. It has been confirmed that cinematographic works and visual media have both positive and negative influences on viewers' stereotypes. Smith et al. (2016) reported various statistics on gender and race/ethnicity percentages for 30,835 people. These people were involved in the production of the annual top-100 films released between 2007 and 2014 (excluding 2011). They reported that only 17 of the top-100 films in 2014 had non-White actors in lead roles. For all 700 films, non-Caucasian actors rarely played stereotypical roles, and there was no change over time.

In a recent study, Besana et al. (2019) analyzed the role and representation of Asian characters in films produced after 1993. They found that Asian characters tend to appear in action/comedy films, and the percentage of Asian characters playing the main roles has suddenly increased since 2016. Ramakrishna et al. (2017) conducted a social network analysis by tabulating the number of contacts between characters from 945 film scripts. Calculating the degree centrality and betweenness centrality of the network for each race, they found that the median betweenness centrality for White characters was significantly higher than that for other races whereas that of Native Americans was significantly lower than those for White, African, and Mixed-race characters. Kagan et al. (2020) collected the subtitles of the most popular feature films on the Internet Movie Database (IMDb) and conducted a social network analysis of the characters, as by Ramakrishna et al. (2017). The PageRank of each character was calculated and compared based on the gender and genre of the characters; there is a higher tendency for male characters to play leading roles in films than female characters, and the number of films in which female characters play leading roles is increasing.

Most of the studies on racial stereotypes and gender bias in the film industry are comparative analyses based on simple statistics of actors in the film industry or are conducted with advanced analytical methods using a small amount of data. In recent years, however, methods for comparative analysis using the social network analysis of numerous film works have been developed, and we now have an environment suitable for the comprehensive analysis of changes in racial representation in the film industry and to elaborate on the trends. Therefore, this study aims to clarify the changes in the roles of characters in the film works from a quantitative perspective by targeting film works that influence viewers' stereotypes, especially Hollywood films that have a global market.

Overview of analysis methods

To clarify the changes in roles of the characters in the film works, the following steps were taken to conduct the analysis:

1. We selected the top 100 films that received the highest ratings for each of the 21 genres of feature films listed on the IMDb. As some films were ranked in multiple genres, 1,217 films were used for the final analysis. We also collected the subtitle data for each film from OpenSubtitle.
2. Based on the studies of Kagan et al. (2020), we created a database of the names of characters, gender, and race of 1,217 movies. Thereafter, the names of characters appearing in the subtitle data and time of speech of the dialogues were extracted.
3. In the subtitles of a particular movie, if any character mentioned another character or if another character appeared within 60 seconds of the playback time of a character's dialogue, it was judged that there was contact between the characters, and the co-occurrence relations of those characters were extracted.
4. We constructed a co-occurrence network of characters from the co-occurrence relations for each movie and calculated the centrality values of each character.
5. We aggregated the values of each centrality by race and compared the changes in the distribution of each centrality over time. Furthermore, the Kruskal-Wallis test and Dunn's test were used to examine whether there was a difference in the median of centrality among the different year groups.

Results and Discussion

Owing to the limited number of words, we focus on the results of changes in degree centrality and betweenness centrality over time. Degree centrality is a simple aggregation of co-occurrence relations. It is an indicator of the frequency of contact between characters in a movie. Betweenness centrality is an index that quantifies the role

of relay points in measuring the shortest distance between nodes (in this case, the characters) for each node. It is an indicator of the character who plays an intermediary function in film work. Figure 1 is an excerpt of a boxplot comparing the values of racial betweenness centrality for different years. The dots in the figure correspond to the centrality values of each actor or character.

Accordingly, we were able to extract the following trends as findings beyond those reported in previous studies:

- From the change in degree centrality, the number of simple contacts between characters was stable and high for White characters throughout the entire period, followed by African and Mixed characters since the 1970s. However, the centrality of other races has increased significantly since the 1980s, and the number of East Asian characters has been increasing since the late 1990s.
- The changes in the betweenness centrality showed that the racial group that played a mediating role was assigned to races other than African in the 1980s. In recent years, East Asians are more likely to appear in films as characters who play a mediating role than Africans.

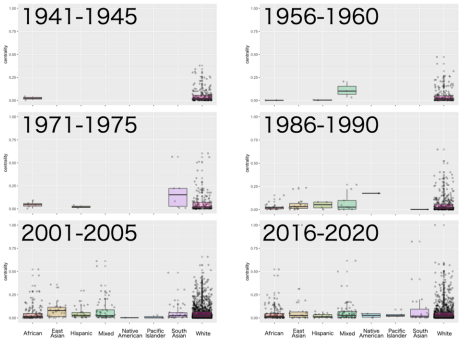


Figure 1:
Distribution of betweenness centrality values of characters by race at 5-year intervals

Bibliography

Behm-Morawitz, E., and Mastro, D. E. (2008). Mean girls? The influence of gender portrayals in teen movies on emerging adults' gender-based attitudes and beliefs, *Journalism and Mass Communication Quarterly*, 85: 131-146.

Smith, S. L., Choueiti, M., Pieper, K., Gillig, T., Lee, C., and Deluca, D. (2016). Inequality in 700 Popular Films: Examining Portrayals of Gender, Race, and LGBT Status from 2007 to 2014, *Institute for Diversity and Empowerment at Annenderg*, 1-22.

Besana, T., Katsiaficas, D., and Loyd, A. B. (2019). Asian American media representation: A film analysis and implications for identity development, *Research in Human Development*, 16: 201-225.

Ramakrishna, A., Martinez, V. R., Malandrakis, N., Singla, K., and Narayanan, S. (2017). Linguistic analysis of differences in portrayal of movie characters, *Proceeding of the 55th Annual Meeting of the Association for Computational Linguistics*, 1669-1678.

Kagan, D., Chesney, T., and Fire, M. (2020). Using data science to understand the film industry's gender gap, *Palgrave Communications*, 6(92): 1-16.

Quantitative Analysis of Gendered Assumptions in a Nineteenth-Century Women's Encyclopedia

Ketzan, Erik
ketzane@tcd.ie
Trinity College Dublin

Hagen, Thora
thora.hagen@uni-wuerzburg.de
University of Würzburg

Jannidis, Fotis
fotis.jannidis@uni-wuerzburg.de
University of Würzburg

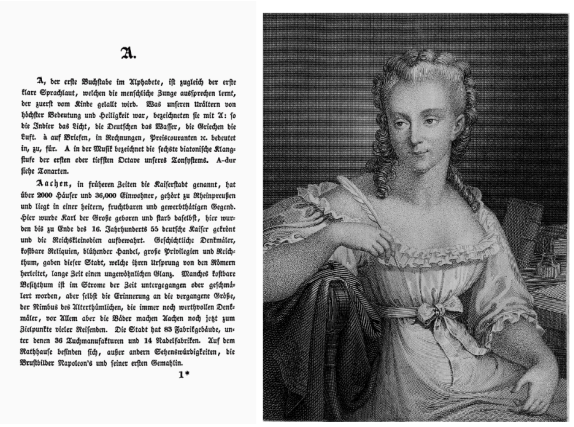
Witt, Andreas
awitt6@uni-koeln.de
IDS Mannheim & University of Cologne

Introduction

This paper quantifies textual patterns relating to gendered assumptions in a fairly unique text, an entire "women's encyclopedia" from 1830's Germany, which at 10 volumes and 1,461,000 word tokens was of comparable size to contemporary general encyclopedias, but written and marketed for a female audience. We perform experiments on classifying gender of biographical entries and querying a specific textual feature, calendar dates, with context from comparison 19th-20th century encyclopedias from the EncycNet corpus.[1]

Encyclopedias in the European tradition were "an element of culture and peoples' lives" (Loveland, 2019), but while encyclopedias invite interpretation (Einbinder, 1966), their length challenges non-digital scholarship; digital humanities thus "holds promise for the study of encyclopedias" (Loveland, 2019).

The *Damen Conversations Lexikon* ("Ladies' Conversations Encyclopedia," hereafter *DamenLex*) has been the subject of little scholarly analysis. Roßbach (2015) writes that the first edition of the *DamenLex* "primarily aims to act as a behavioral guide for virtuous women,"[2] but Schaser (2016) asserts that this "little known" encyclopedia is "a treasure trove for questions of cultural history." *DamenLex* editor Carl Herloßsohn (1804-1849) explained its content selection in gendered assumptions and value judgments, and the extent to which the *DamenLex* actually followed Herloßsohn's stated goals is the starting point of our research questions.



A.

A, der erste Buchstabe im Alphabete, ist zugleich der erste klare Sprachlaut, welchen die menschliche Zunge aussprechen lernt, der zuerst vom Kinde gelallt wird. Was unseren Urältern von höchster Bedeutung und Heiligkeit war, bezeichneten sie mit A: so die Indier das Licht, die Deutschen das Wasser, die Griechen die Luft. à auf Briefen, in Rechnungen, Preiscouranten ꝛc. bedeutet in, zu, für. A in der Musik bezeichnet die sechste diatonische Klangstufe der ersten oder tiefsten Octave unseres Tonsystems. A-dur siehe Tonarten.

Aachen, in früheren Zeiten die Kaiserstadt genannt, hat über 2000 Häuser und 36,000 Einwohner, gehört zu Rheinpreußen und liegt in einer heitern, fruchtbaren und gewerbthätigen Gegend. Hier wurde Karl der Große geboren und starb daselbst, hier wurden bis zu Ende des 16. Jahrhunderts 55 deutsche Kaiser gekrönt und die Reichskleinodien aufbewahrt. Geschichtliche Denkmäler, kostbare Reliquien, blühender Handel, große Privilegien und Reichthum, gaben dieser Stadt, welche ihren Ursprung von den Römern herleitet, lange Zeit einen ungewöhnlichen Glanz. Manches kostbare Besitzthum ist im Strome der Zeit untergegangen oder geschmälert worden, aber selbst die Erinnerung an die vergangene Größe, der Nimbus des Alterthümlichen, die immer noch werthvollen Denkmäler, vor Allem aber die Bäder machen Aachen noch jetzt zum Zielpunkte vieler Reisenden. Die Stadt hat 83 Fabrikgebäude, unter denen 36 Tuchmanufakturen und 14 Nadelfabriken. Auf dem Rathhause befinden sich, außer andern Sehenswürdigkeiten, die Brustbilder Napoleon's und seiner ersten Gemahlin.

1*

Figure 1.
A sample page and illustration from the Damen Conversations Lexikon

Related work

Distant reading the *DamenLex* for ideological traces relates to issues of women and gender in 19th-century Europe, women's education and achievement, and the history of books not **by**, but marketed **for** women, in which women's access to the written word, by controlling literacy and access to reading material, has been a source of anxiety (Jack, 2012). We would thus expect the *DamenLex* to display evidence of two opposing forces: women's education and controlling ideology through explicit or implicit gender presumptions within the content, stylistics, and selection of topics.

Perceptions of women readers can be traced through such texts as Ovid's books addressed to women (e.g. *Ars Amatoria*) and the Confucian *Four Books for Women* (Mingqi, 1987). Eighteenth- and nineteenth-century texts for women on etiquette and conduct were prescriptive, supporting notions of "ideal womanhood" (Hemlow, 1960; Darby, 2000). The nineteenth century in which the *DamenLex* was published was "a golden age for reading, and for women's reading in particular," per Jack (2012), as the growth of industrialization, printing and publishing were "accompanied by wide-ranging debates about what women [...] should be encouraged to read, or discouraged — even prevented — from reading."

Gendered assumptions

Gender bias in biographical entry subjects is an unfortunate theme in the history of the encyclopedia, with the 11th Britannica, for instance, including an entry on Pierre Curie but not Marie Curie (Thomas, 1992), and Bamman and Smith (2014) estimated that only 14.8% of biographical entries in Wikipedia have women as subjects.

The *DamenLex*'s editor explicitly included many biographies of women to appeal to women readers, suggesting experiments to classify the gender of the ~800 biographical entries in the *DamenLex* and almost 44,000 in comparison encyclopedias.[3] To classify biographical entries, we trained a bag-of-words-based SVM classifier to label entries as either biography, place, object or abstract concept (with an accuracy of 0.92 for biographies). To classify the gender of each biographical entry, a rule-based approach based on Reagle and Rhue (2011) compares the ratios of male and female personal pronouns in the entry (e.g. sein/his, ihr/her). Only entries longer than 20 tokens were classified, and we only proceeded with entries for which a gender was identified. The amount of unclassified entries is low for *Brockhaus 1837* (about 1%), where entries tend to be relatively long, but higher in *Brockhaus 1911* (about 37%), for example.

	Brockhaus 1809	DamenLex 1834	Brockhaus 1837	Herder 1854	Meyer 1905	Brockhaus 1911	Wikipedia 2014	Wikipedia 2015
Male	948 (95%)	480 (60%)	957 (96%)	5,952 (95%)	26,223 (94%)	7,356 (94%)	85.20%	84.50%

Female	52 (5%)	329 (40%)	42 (4%)	345 (5%)	1,599 (6%)	479 (6%)	14.80%	15.50%

Table 1. Estimate of male and female biographies in historical German encyclopedias. Wikipedia results reported by Bamman and Schmith (2014) and Graells-Garrido et al. (2015)

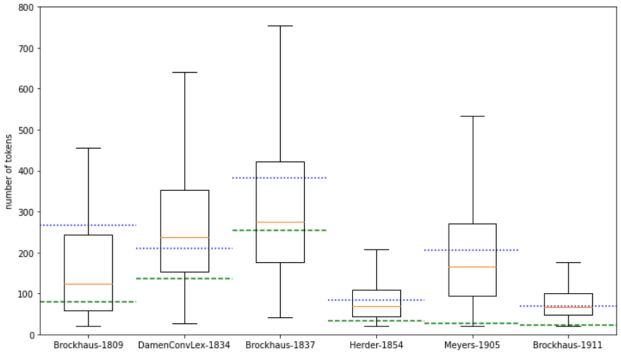


Figure 2.
Boxplot visualizing the entry lengths of female biographies in tokens in all encyclopedias. The solid, orange line marks the median. Additionally, the median of male biography entry lengths (dotted, blue line) and the median of all entry lengths, including biographies, (dashed, green line) are given

From these results, *DamenLex* contains a much higher percentage of female biographies than all other comparison texts (Table 1) including Wikipedia, with female biographies around 40%. Two chi-squared tests (Table 4 in the appendix) reveal that there is a highly significant relationship between encyclopedia and the amount of entries on women only when the *DamenLex* is included in the test, confirming our hypothesis. A similar gender disparity is observed in entry lengths of of male and female biographies in the *DamenLex* (Figure 2). We calculated Mann-Whitney U tests for article lengths of female biographies compared to similar sized samples from male biographies for all encyclopedias (see Table 3 in the appendix). Only the *DamenLex*'s and *Brockhaus' 1911* lengths of female biographies are not significantly shorter than the male entries.

Did *DamenLex* devote biographies to the same "notable" women as contemporary encyclopedias? The overlap is quite low: of 329 entries on women in *DamenLex*, only about 65 appear in at least one other encyclopedia, confirming that the editors of the comparison encyclopedias had a different perception of who "important women" were.

Finally, most frequent words in *DamenLex* biographical entries provide insight into content differences. Among the 15 most frequent nouns in female biographies are role labels such "daughter," "wife," and "mother," and family relations such as "husband" and "child." In male biographies, in contrast, only "father" and "son" appear in the 20 most frequent nouns, while references to artistic production such as "poet," "poem," "opera," and "writing" fill out the rest. Such artistic terms can also be found in female biographies, only at lower frequency ranks. Among the 50 most frequent adjectives in male entries, only a handful do not appear in the 100 most frequent adjectives in female entries: "tremendous," "glowing," "musical," and "exquisite." The first two words are typical terms to describe sublime aesthetic experiences, which aligns with contemporary gendered assumptions about aesthetics.

Herloßsohn wrote that a "romantic representation" of the subjects in the *DamenLex* was desired: "not a tiresome enumeration of facts and the course of time, but a lively, rapidly gliding painting [...] should be given." We thus hypothesize that the amount of calendar dates will be far lower in its entries. To investigate, we tagged encyclopedia entries with heideltime,[4] a multilingual temponym tagger. As Table 2 shows, the amount of dates is indeed far lower in *DamenLex* than comparison encyclopedias, confirming the gendered assumption that hard facts such as calendar dates were considered undesirable for women readers. Verification via the chi-squared test (Table 4 in the appendix) results that there is no statistical difference between the encyclopedias concerning the amount of dates, however.

	Brockhaus 1809	DamenLex 1834	Brockhaus 1837	Herder 1854	Meyer 1905	Brockhaus 1911
Dates	1.29%	0.73%	1.38%	1.87%	3.81%	4.32%

Table 2. Relative amount of dates found in a sample of 256 entries of similar length (about 100,000 tokens overall) per encyclopedia

Conclusion

By quantifying the ratios of male/female biographical entries in the *DamenLex* and comparison encyclopedias, comparative length of biographical entries, and a query of calendar dates in the texts, we provide new knowledge and add historical context to vibrant ongoing research on gender bias in encyclopedias (including Wikipedia). We agree with Schaser (2006) that the "little known" encyclopedia of the *DamenLex* is "a treasure trove for questions of cultural history," and have presented evidence that distant reading of gender distribution in biographical entries and content

presentation can reveal gendered assumptions in the text. This paper will include these and other experiments to quantify gendered assumptions in encyclopedia texts, and could support future work in gender bias in not only historical but also contemporary encyclopedias.

Appendix

Brockhaus 1809	$U = 737.0$, $n1 = n2 = 52$, $p < .001$
Brockhaus 1837	$U = 758.5$, $n1 = n2 = 42$, $p < .001$
Brockhaus 1911	$U = 112124.5$, $n1 = n2 = 479$, $p = 0.27$
Herder 1854	$U = 49367.0$, $n1 = n2 = 345$, $p < .001$
Meyer 1905	$U = 728612.0$, $n1 = n2 = 1599$, $p < .001$
DamenLex 1834	$U = 57677.5$, $n1 = n2 = 329$, $p = 0.93$

Table 3. Results of the one-sided Mann-Whitney U tests ($p<.01$) to confirm or reject the hypothesis whether male biography entries are significantly longer than female biography entries per encyclopedia

Number of entries (incl. DamenLex)	$\chi2$ (5, $N = 44762$) = 1635.9, $p < .001$
Number of entries (excl. DamenLex)	$\chi2$ (4, $N = 43953$) = 7.7, $p = .1$
Number of dates	$\chi2$ (5, $N = 600$) = 5.03, $p = .41$

Table 4. Results for the chi-squared tests ($p<.01$) for the amount of entries on men and women as well as the amount of dates (and non-dates, in percent) per encyclopedia. For the dates, we opted for choosing the percentages over the raw counts, as the sample size makes the interpretation of the otherwise very low p-values difficult

Bibliography

Bamman, D. and Smith, N. A. (2014). Unsupervised Discovery of Biographical Structure from Text. *Transactions of the Association for Computational Linguistics*, 2: 363-76.

Darby, B. (2000). The more things change… the rules and late Eighteenth-Century Conduct Books for Women. *Women's Studies*, 29(3).

Einbinder, H. (1966). The Myth of the Britannica. *Science and Society*, 30(1).

Graells-Garrido, E., Lalmas, M. and Menczer, F. (2015), First Women, Second Sex: Gender Bias in Wikipedia. *Proceedings of the 26 th ACM Conference on Hypertext & Social Media*, pp. 165-74.

Hagen, T., Ketzan, E., Jannidis, F. and Witt, A. (2020). Twenty-two Historical Encyclopedias encoded in TEI: a new Resource for the Digital Humanities. *Proceedings of the 4 th joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pp. 112-20.

Hemlow, J. (1960). Fanny Burney and the Courtesy Books. *PLMA*, 65(6).

Jack, B. (2012). *The Woman Reader.* New Haven: Yale University Press.

Loveland, J. (2019). *The European Encyclopedia: From 1650 to the Twenty-first century*. Cambridge: Cambridge University Press.

Mingqi, Z. (1987). The Four Books of Women: Ancient Chinese Texts for the Education of Women. *B.C. Asian Review*, 1(1).

Reagle, J. and Rhue, L. (2011). Gender Bias in Wikipedia and Britannica. *International Journal of Communication*, 5(0).

Roßbach, N. (2015). *Wissen, Medium und Geschlecht: Frauenzimmer-Studien zu Lexikographie, Lehrdichtung und Zeitschrift.* Frankfurt am Main: Peter Lang International Academic Publishers.

Schaser, A. (2006). Rezension zu: Herloßsohn, Carl (Hrsg.): Damen Conversations Lexikon. Neusatz und Faksimile der 10-bändigen Ausgabe Leipzig 1834 bis 1838. Berlin 2005. *H-Soz-Kult.*

Thomas, G. (1992). *A Position to Command Respect. Woman and the Eleventh Britannica*. Metuchen, NJ: Scarecrow Press.

Notes

1. https://encycnet.github.io/
2. Translation of quotations by Roßbach (2015) and Schaser (2016) in this paragraph are by us.
3. The encyclopedias for our experiments are part of a larger set of historical reference works converted to TEI (Hagen et al., 2020): http://dx.doi.org/10.5281/zenodo.4039569.
4. https://github.com/HeidelTime/heideltime

On Digitizing Historic Music Storage Media For Computational Analysis

Khulusi, Richard
khulusi@informatik.uni-leipzig.de
Image and Signal Processing Group, Computer Science, Leipzig University, Germany

Fricke, Heike
heike_marianne.fricke@uni-leipzig.de
Research Center Digital Organology, Musical Instrument Museum, Leipzig University

Fuhry, David
david.fuhry@uni-leipzig.de
Research Center Digital Organology, Musical Instrument Museum, Leipzig University

Piontkowitz, Vera
vera.piontkowitz@uni-leipzig.de
Research Center Digital Organology, Musical Instrument Museum, Leipzig University

Focht, Josef
Josef.Focht@uni-leipzig.de
Research Center Digital Organology, Musical Instrument Museum, Leipzig University

Introduction

Between 2018 and 2020, the digitization project TASTEN, funded by the German government, digitized 3200 piano rolls for self-playing pianos preserved at the Musical Instruments Museum at Leipzig University (MIMUL). A piano roll is a historic music storage media (MSM) coding movement impulses through holes punched into paper (Focht, 2020). When played, a pneumatic system uses this code to create sounds at runtime. For musicologists, such piano rolls are of high value. First, they are the only source of musical performances by famous pianists in times preceding sound recording technologies. Secondly, for researchers they offer a rich repository to study the music-making practice and musical interpretation around 1900. We expand our prior source catalogue to include their technical predecessors: Cardboard and metal plates, which also use punched holes as code (Focht, 2021). They share the issue of being fragile – after decades of usage and storage – making it important to digitize them for preservation, also.

Related Work

We found a lack of publications dealing with challenges, issues, and processing of MSM in general. This may be due to most of such projects being on a small scale in private settings (pianola.co.nz, 2021; IAMMP, 2021). Scientific publications are presented by Debrunner (2013), who developed a scanner capable of reading information directly from piano rolls and dealing with paper based distortion issues, for a single format (Welte piano rolls). Shi et al. (2019) built an online database of almost 500 digitized piano rolls offering representations including images, audios, and MIDI files, focusing only on the same format. No published scientific or private projects are to be found concerning metal or cardboard plates and their digitization.

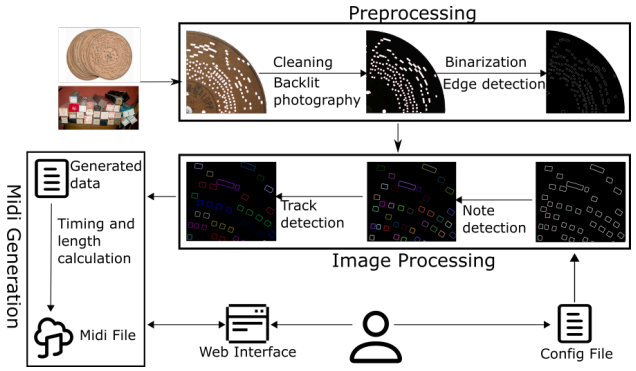


Figure 1:
Pipeline of the digitizing of piano rolls and plates through different processing steps and manual input of a Musicologist

Digitizing MSM

While prior projects show great results in digitizing piano rolls, processes capable of dealing with multiple formats of piano rolls and plates in general are non-existent. Furthermore, digitizing is only the first step for musicologists. This data allows them to answer research questions by reading and hearing the encoded works which are difficult to impossible to read. Additionally, these processes allow for distant reading analysis and comparative approaches. We propose a workflow (Figure~1) capable of digitizing all 3,200 piano rolls of 30 different formats available in the MIMUL. Currently, work is done to include more than 25 formats of 438 plates. The workflow begins with a conservator-led cleaning process to protect

the objects and the researchers. Damages which would lead to the destruction of the object were documented. Next, metadata like weight, measurements, format, title, composer, and performer was extracted to be included in our research tool musiXplora. For the actual digitizing, a scanning company was commissioned, for which we constructed an unwinding mechanism, making it possible to create a single scan of the piano rolls (as 300dpi .tif images, resulting in up to 5.000x550.000 pixels and up to 5GB). While the prior project (TASTEN – 2018-2020) generated scans of piano rolls, the current DISKOS project focuses more deeply on musicologists' research questions. Examples would be "How did composers play their own compositions on the piano?", or "Can the computer use this digital knowledge to identify which pianist played a piano roll of unknown origin?" Also, exploration and visual analysis of the objects can be offered through distant reading visualization systems embedded in the musiXplora.

Technical Details

Starting with the preprocessing of the backlit images the actual process labels connected components (musical notes) on the image and applies filtering to keep only the components representing notes. Calculating the distance from each hole to the edge of the medium and using mean shift clustering we can assign each note to its respective track. We finally apply corrections to account for empty tracks and distortions and calculate the position and width of each hole. Combining this information with an expert created mapping of tracks to MIDI notes, we can then generate a MIDI file accurately representing the information on the medium. Relevance for Musicology These files allow users to work interactively with the music and help musicologists in their work with these sources. Furthermore, this offers an enhanced experience for museum visitors, by making it possible to voice the digitized piano rolls and plates using the digital representation of keyboard instruments created during TASTEN. Hence, historic media can be experienced on historic instruments even if the physical instruments would not have been interoperable. For the musicologists, these results offer a way to open up previously unreadable sources. Besides close reading approaches, the generated data allows for distant reading (visualization) methodology and novel research questions like examining which schools of interpretation and playing techniques are represented and how they have spread between performers. Further, these results are also important for educational and playful aspects like listening to these historical virtuous interpretations without risking the media and instruments.

Limitations

As we are in the first of three project years, some challenges still exist: The image processing results are highly dependable on a suitable preprocessing, which differs quite a lot even between different formats of the same media type. Metal plates in particular are operated under pressure during playback. While the playback instruments themselves are constructed to negate this deformation, a plain photography of the media does not lead to correct results and needs a semi-manual correction process. In general, we are content with our results for cardboard plates and piano rolls of specific formats, but are aware of needed improvements.

Conclusion

Even with the best preservation techniques, most materials deteriorate from humidity and stress through time. Hence, valuable information stored on analog media is prone to damage and even loss. Such information includes original recordings of famous virtuosos like Edvard Grieg, not only valuable as music recording, but also important for musicologists interested in analyzing playing techniques and differences between musical notation and interpretation by their composers. To allow digital processing of such media, digitization is mandatory. We present a pipeline taking image sources of circular and linear played storage media and creating digital representations in form of MIDI files, which then can be analyzed, further processed, edited, or even played through modern and historical instruments.

Bibliography

pianola.co.nz(2021): Saving the Music of Yesterday. *http://www.pianola.co.nz/public/index.php* (Accessed: 01 January 2022)

Comaniciu, D., & Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5), 603–619. https://doi.org/10.1109/34.1000236

Debrunner, D. (2013). Von der Welte-Rolle zu parametrisierbaren Wiedergabe auf synthetischen Instrumenten und MIDI-fähigen Selbstspielklavieren. *In C. E. Hänggi & Köpp (Eds.), Recording the Soul of Music: Welte-Künstlerrollen für Orgel und Klavier als authentische Interpretationsdokumente.*

Focht, J. (2020). *MusiXplora*. https://musixplora.de/mxp/2002522 (Accessed 01 January 2022)

Focht, J. (2021). *MusiXplora.* https://musixplora.de/mxp/2003518 (Accessed 01 January 2022)

IAMMP(2021). *International Association of Mechanical Music Preservationalists*. http://www.iammp.org/ (Accessed 01 January 2022)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Shi, Z., Sapp, C. S., Arul, K., McBride, J., & Smith III, J. O. (2019). Supra: Digitizing the Stanford University Piano Roll Archive. *In A. Flexer, G. Peeters, J. Urbano, & A. Volk (Chairs), ISMIR, Delft*.

van der Walt, S., L. Schönberger, J., Nunez-Iglesias, J., Boulogne, F., D. Warner, J., Yager, N., Gouillart, E., Yu, T. (2014). scikit-image: Image processing in Python.*PeerJ* 2:e453 https://doi.org/10.7717/peerj.453

The ImageMagick Development Team.(2021). *ImageMagick*. Retrieved from https://imagemagick.org (Accessed 01 January 2022)

Book Barcoding: A Framework for the Visual Collation and Woodblock Tracking of Japanese Printed Books

Kitamoto, Asanobu
kitamoto@nii.ac.jp
ROIS-DS Center for Open Data in the Humanities, Japan; National Institute of Informatics

Introduction

Printed books have played a central role in the distribution of knowledge. In the publishing industry during the Japanese Edo period (1603-1868), woodblock printing became more popular than movable type printing because it fits better with the features of the Japanese language. However, because woodblock was very expensive to create from scratch, information updates were usually applied as a patch to the woodblock. Hence the visual overlay and comparison of two images from the same woodblock printed at different times can reveal small changes between the two versions. Therefore, technology for keeping track of the same woodblock over time is critical in answering research questions such as which part or how often information was updated on the woodblock over time.

Our proposed algorithm, 'woodblock tracking' and 'visual collation,' solves this problem. First, woodblock tracking compares two books to identify a pair of pages that originate in the same woodblock. Second, visual collation compares a pair of pages to highlight pixel-level differences by estimating a projective transformation matrix to overlay those images. The proposed algorithm was applied to the comprehensive analysis of the Bukan Complete Collection < http://codh.rois.ac.jp/bukan/ >, which contains 381 Bukan book series published for nearly 200 years in the Edo period (Kitamoto, 2018). The algorithm automatically identifies the series of images originating in the same woodblock, and we use this result for 'differential transcription' to realize efficient transcription over the book collection that changes seamlessly over time.

Algorithm

Visual collation

In comparison to textual collation, visual collation has a few advantages. First, visual collation does not require costly transcription. Second, non-textual content, such as graphic elements and physical changes on the woodblock, can be considered. However, traditional visual collation requires playing a 'spot-the-difference game' through manual side-by-side comparison, which is time-consuming and highly unreliable. Hence we proposed a computer vision-based algorithm to automate this process (Leyh, 2020).

The algorithm extracts keypoints from images with descriptors associated with each keypoint. Those keypoints are then used for image comparison by computing the distance between descriptors of matched keypoints. Finally, the algorithm computes a projective transformation matrix to overlay two images based on inlier keypoints. Here the number of inlier keypoints roughly represents matching quality. Because the algorithm used is standard in computer vision, we used the off-the-shelf library OpenCV, as shown in Figure 1, by searching for the best settings from available algorithms. We also developed a web-based image comparison tool, 'vdiff.js' < http://codh.rois.ac.jp/software/vdiffjs/ >, to interact with visual collation on a web platform. This tool allows four comparison modes: slider, emphasis, red-blue, and side-by-side, and is helpful for 'differential reading' to focus on information changed between two images.

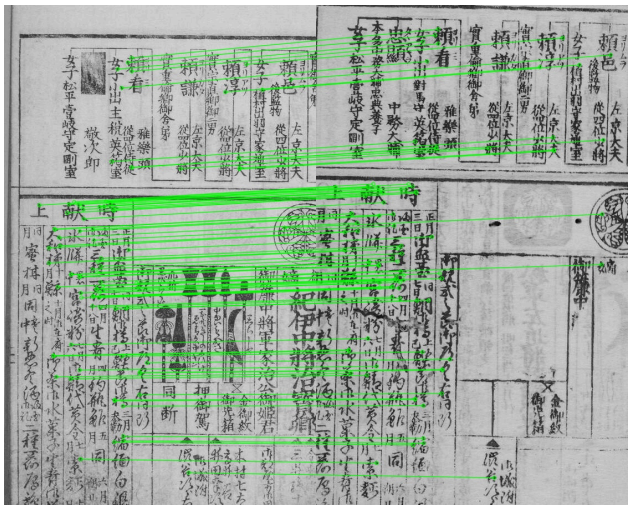


Figure 1:
Keypoint matching of two images. We used the AKAZE feature detector (Alicantarilla, 2011), Hamming distance as the distance metric between descriptors, and the RANSAC algorithm (Fischler, 1981) to compute a projective transformation matrix.

Woodblock tracking

Woodblock tracking uses the result of visual collation to identify a set of images printed by the same woodblock. We formulated this problem as connecting the path of best-matching image pairs across books. First, we search for best-matching image pairs between two books based on the number of inlier keypoints as the score of matching quality. For this purpose, we use the Gale-Shapley (GS) algorithm (Gale, 1962), which is a classic algorithm to solve a stable marriage problem in operations research, to find the best match between two books. We then extend the matching from two books to multiple books. For example, having the matching results for Book A and B, we can either extend the path to the neighboring books, such as Book C and A and Book B and D, to form a path C-A-B-D. After obtaining all the paths, we assign a unique ID for each woodblock and keep track of changes that occurred in the same woodblock.

Results

We applied the proposed algorithm to the Bukan Complete Collection, a dataset of 381 Bukan books released from ROIS-DS Open Data Center for Humanities (CODH), derived from digitized images from the National Institute of Japanese Literature (NIJL). Bukan is a book about the directory of families of the state king (Daimyo) and bureaucrats of the central government (Bakufu) in the Edo period. It had been a best seller book for as long as 200 years. Moreover, it had been updated and published frequently with a peak frequency of a few times in a month. We believe that those different versions of Bukan are unique historical materials to reconstruct the time-series of biographical information of the period.

First, we selected 354 books relevant for the analysis, which amounts to 150,698 images. Second, we tried matching on 85,990,142 image pairs and selected 541,810 image pairs for visual collation. Third, we applied woodblock tracking to identify 44,842 woodblocks and found that the longest life was across 49 books. This result demonstrates that the proposed algorithm can scale to a level that manual annotation can never achieve. All results are available on the Bukan differential reading platform < http://codh.rois.ac.jp/bukan/diff/ >, and we pick up some results from Figures 2 through 4.

We finally used this platform for differential transcription. We focused on one Daimyo family and transcribed a few attributes of the family. Among 354 books, information about the family was found in 203 books, among which 165 books were collated by woodblock tracking, while 38 books were not collated due to different woodblock layouts. This result is beneficial for efficient time-series transcription while giving us hints about how often the woodblock was updated.



Figure 2:
The list of visual collations computed between two books. Background color from red to green, blue represents the ascending order of the score, computed from the number of inlier keypoints, while the gray color represents a page without a corresponding page.

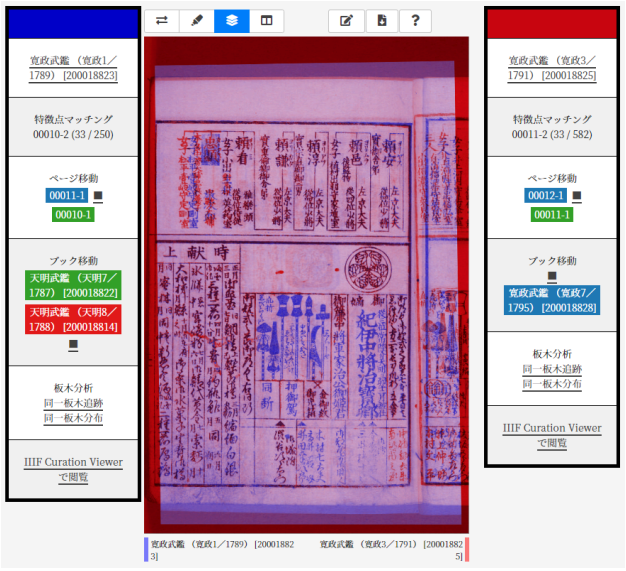


Figure 3:
The result of the page-by-page collation visualized as the superimposition of two images using vdiff.js. The blue color shows pixels from the book on the left, while red on the right.



Figure 4:
The result of woodblock tracking. The vertical axis represents the book's page number starting from the top, and the horizontal axis represents the order of books by estimated publication dates. A long horizontal line indicates a woodblock that survived longer, and the red line highlights the woodblock with the longest life.

Book Barcoding

Finally, we named the proposed algorithm ‘book barcoding,’ whose name was inspired by ‘DNA barcoding’ (Moritz, 2004). A DNA barcode is a DNA sequence that is specific to a species. When investigating the species of a particular DNA sequence, the sequence is compared with the DNA barcode library, such as BOLD (Barcode of Life Data Systems), to identify DNA sequences from unknown species. Based on a similar framework, we plan to establish a general collation platform for printed books, where keypoints specific to a book helps to identify the phylogenetic relationship of unknown books.

Acknowledgment

The author thanks Mr. Jun Homma for his significant contribution to vdiff.js. He also thanks Prof. Kumiko Fujizane and Prof. Kazuaki Yamamoto of the National Institute of Japanese Literature for helpful comments on the research. A part of the research is based on the work of Mr. Thomas Leyh, who contributed to this project while he was an NII internship student. JSPS KAKENHI Grant Number JP19H01141 partially supports this work.

Bibliography

Alcantarilla, P. F., Nuevo, J., and Bartoli, A. (2011) Fast explicit diffusion for accelerated features in nonlinear scale spaces. Trans. Pattern Anal. Machine Intell, 34:7, 1281–1298.

Fischler, M. A., and Bolles, R. C. (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24:6, 381–395.

Gale, D. and Shapley, L. S. (1962) College Admissions and the Stability of Marriage. The American Mathematical Monthly, 69:1, 9-15.

Kitamoto, A., Horii, H., Horii, M., Suzuki, C., Yamamoto, K., Fujizane, K. (2018) Differential Reading by Image-based Change Detection and Prospect for Human-Machine Collaboration for Differential Transcription, Digital Humanities Conference.

Leyh, T., Kitamoto, A. (2020) Computer Vision-based Comparison of Woodblock-printed Books and its Application to Japanese Pre-modern Text, Bukan. Tenth Conference of Japanese Association for Digital Humanities (JADH2020), 53-59.

Moritz, C., Cicero, C. (2004) DNA Barcoding: Promise and Pitfalls. PLoS Biol 2:10, e354.

Emotions and Literary Periods

Konle, Leonard
leonard.konle@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Germany

Kröncke, Merten

merten.kroencke@uni-goettingen.de
Georg-August-Universität Göttingen, Germany

Jannidis, Fotis
fotis.jannidis@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Germany

Winko, Simone
simone.winko@phil.uni-goettingen.de
Georg-August-Universität Göttingen, Germany

> "Longing, resignation, derision, disillusionment, weary smiles, these are the five basic tones of the modern scale of emotions." (Servaes 1896)

Introduction

Periodization has been neglected in computational literary studies despite some early discussions (Underwood 2013). In literary studies the usual basis for the construction of periods are differences in the choice of topics or style or non-literary aspects, while differences in the representation of emotions are underresearched. This is the case even though recent approaches in literary studies ascribe epoch-specific relevance to the literary representation of emotions. How to use quantitative methods to study emotions in literary texts and use them to describe the differences between periods is the focus of our paper; our use case is the difference between realism and early modernism in German literary history and we are focusing on poetry. In a first step a group of domain experts manually annotated around 1.000 poems, highlighting phrases according to the emotions they represented. In the second step a machine learning model was trained and in a third step this model was used to annotate a collection of more than 6.000 poems, from anthologies representing either realism or early modernism. Lastly we analyzed the main differences of these periods based on the trends we found. [1]

Resources

Our corpus [2] consists of 6249 poems from 20 anthologies. 12 anthologies, published between 1885 and 1911, are explicitly intended by the editors to contain 'modern' poetry. [3] The other anthologies were published between 1859 and 1882 and represent the earlier poetry of realism.

We gathered emotion annotations for 1278 poems. The goal was not to annotate readers' emotions, but rather the emotions represented in the text itself. The annotators used a list of 40 discrete emotions (see Table 1), the selection of which was based both on existing emotion models (e.g. Ekman 1992, 1999; Plutchik 1980a, 1980b, 2001) and on the emotions that were regularly represented in the poems of our corpus. We categorized the emotions into 6 groups, inspired by the emotion hierarchy in (Shaver et al. 1987). First, each poem was annotated independently by two annotators, then they merged annotations manually into a consensus annotation. Their agreement, measured with γ (Mahet et al. 2015), was 0.6445 for individual emotions and 0.7491 for the emotion groups.

Table 1: Emotions and Emotion Groups

Love	Joy	Surprise/ Agitation	Anger	Sadness	Fear
Admiration	Bliss	Agitation	Anger	Despair	Fear
Affection	Calmness	Emotionality	Contempt	Disappointment	Fright
Desire (non-sexual)	Comfort	Surprise	Disgust	Impatience	
Gratefulness	Enthusiasm	Suspense	Dislike	Insecurity	
Longing	Hope		Envy	Melancholy	
Love	Joy		Hate	Loneliness	
Lust (sexual)	Pride			Pity	
	Solace			Powerlessness	
				Regret	
				Sadness	
				Shame	
				Suffering	
				Uneasiness	

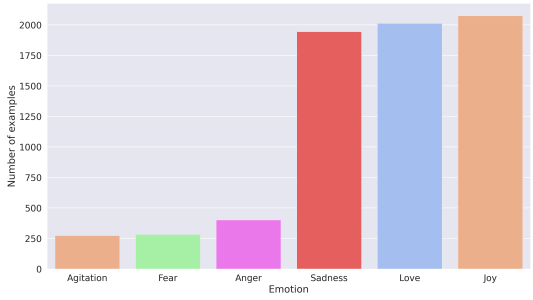


Figure 1:
Provided examples per grouped Emotion.

The emotions groups are not equally balanced (see Fig. 1). This distribution could be specific to our corpus and very probably will change with other genres.

Emotion Classification

We model emotion classification as a series of binary classifications to avoid the complexity of a multi-labeling task. Basis of our classification experiment is the german BERT (Devlin et al. 2018) model gbert-large (Chan et al. 2020). Because gbert is trained on contemporary webtext, we continue its pre-training [4] with poetry to adapt to our target domain. Subsequently we perform fine-tuning on the binary emotion classification tasks. To overcome the class imbalance we apply undersampling by randomly sampling examples from the majority class in every epoch. While the classification of single emotions leads to a large spread in predictive quality [5], the grouped emotions (Table 1) lead to more stable performance at an acceptable level of uncertainty (Table 2).

Table 2: Quality of Emotion Classification.

	Joy	Love	Sadness	Anger	Fear	Agitation
f1 (macro)	0.73	0.77	0.74	0.71	0.79	0.62

Analysis

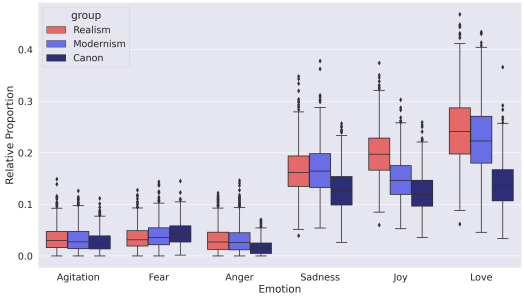


Figure 2:
1000 Samples of 20 poems drawn out of the emotion predictions of each group.

Our results (Fig. 2) show that modernist poetry as a whole represents emotions slightly less frequently than realist poetry, but the effect sizes are small. 9% of realist poems and 12% of modernist poems do not represent any emotion. The probability that a verse contains an emotion is 47% in realism and 42% in modernism. The decrease in emotionality from realism to modernism is mainly due to the emotion group joy, i.e. positive emotions.

If only canonical modernist authors [6] are considered, the tendency to represent fewer emotions is much stronger. The probability that a poem from a canonical author does not represent any emotion is 14%, and the probability that a verse from the canonical subcorpus contains an emotion is 39%. Not only joy, but also anger, sadness, and especially love become less frequent compared to the poetry of realism. Again, the decrease is most pronounced for positive emotions.

Discussion

Some literary scholars claim that German modernist poetry, in contrast to the more traditional poetry of realism, tends toward a sober, matter-of-fact, and non-emotional mode of expression (cf. e.g. Andreotti 2014). Others argue that modernist poetry does indeed represent emotions frequently, albeit in a modified way (cf. e.g. Winko 2003). Our results support the view that modernist poetry as a whole continues to represent emotions frequently, that is, almost as frequently as the poetry of realism. There is a much more significant decrease in emotionality, however, when considering only canonical authors. This suggests that the contradicting views in literary studies regarding the emotionality or non-emotionality of modernist poetry could be explained, at least in part, by different objects of study. The scholars who support the non-emotionality thesis might have focused more than the others on canonical authors. These observations highlight the importance of selection processes and corpus formation in literary history. Future research could examine further selection criteria and categories, such as gender or class.

The trend to represent emotions less frequently applies especially to positive emotions. As a result, negative emotions make up a larger proportion of the remaining emotions and modernist poetry appears more negative overall. This is an interesting topic for further research. Moreover, it seems instructive to investigate later literary periods such as expressionism. In addition, it should be interesting to examine mixed emotions. Finally, it is desirable to not only analyze the frequency of emotions, but also the way of representation, e.g. explicit or implicit modes, which is especially important when dealing with literature.

Bibliography

Andreotti, Mario. Die Struktur der modernen Literatur. Neue Formen und Techniken des Schreibens: Erzählprosa und Lyrik. P. Haupt, 5th edition 2014.

Chan, Brandon; Schweter, Stefan and Möller, Timo. (2020): German's next language model In: Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online), pp. 6788–6796. URL: https://aclanthology.org/2020.coling-main.598. doi: 10.18653/v1/2020.coling-main.598.

Devlin, Jacob; Chang, Ming-Wei; Chang; Kenton, Lee and Toutanova, Kristina (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Ekman, Paul. "An Argument for Basic Emotions." Cognition and Emotion, vol. 6, no. 3-4, 1992, pp. 169–200.

Ekman, Paul. "Basic Emotions." Handbook of Cognition and Emotion, edited by John Tim Dagleish and Mich J. Power. Wiley, 1999, pp. 45-60.

Gururangan, Suchin; Marasović, Ana; Swayamdipta, Swabha; Lo, Kyle; Beltagy, Iz; Downey, Doug; and Smith, Noah A. (2020): Don't stop pretraining: adapt language models to domains and tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

Konle, Leonard and Jannidis, Fotis (2020): Domain and Task Adaptive Pretraining for Language Models. CHR 2020: Workshop on Computational Humanities Research, November 18–20, 2020, Amsterdam, The Netherlands. Proceedings http://ceur-ws. org ISSN, 1613, 0073.

Mathet, Yann; Widlöcher, Antoine; Métivier, Jean-Philippe (2015): The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment Computational Linguistics, MIT Press, September 2015, Vol. 41, No. 3: 437-479.

Plutchik, Robert. Emotion: A Psychoevolutionary Synthesis. Harper & Row 1980a.

Plutchik, Robert. "A general psychoevolutionary theory of emotion." Emotion: Theory, Research and Experience. Theories of Emotion, edited by Robert Plutchik and Henry Kellerman. Academic Press, 1980b, vol. 1, pp. 3–33.

Plutchik, Robert. "The Nature of Emotions." American Scientist, vol. 89, no. 4, 2001, pp. 344–350.

Servaes, Franz. Goethe am Ausgang des Jahrhunderts. In: Neue deutsche Rundschau (1896), pp. 1073-1090 (translation by FJ/SW).

Shaver, Phillip, et al. "Emotion Knowledge: Further Exploration of a Prototype Approach." Journal of Personality and Social Psychology, vol. 52, no. 6, 1987, pp. 1061–1086.

Underwood, Ted: Why Literary Periods Mattered? Stanford University Press 2013.

Winko, Simone. Kodierte Gefühle. Zu einer Poetik der Emotionen in lyrischen und poetologischen Texten um 1900. Erich Schmidt, 2003.

Notes

1. CRediT Roles: Leonard Konle: Inve stigation, Data Curation, Writing – original draft; Merten Kröncke: Data Curation, Writing – original draft; Fotis Jannidis: Conceptualization, Supervision, Writing – review & editing; Simone Winko: Conceptualization, Writing – review & editing.
2. Code and data: https://github.com/LeKonArD/Emotions-and-Literary-Periods
   Corpus Release: https://doi.org/10.5281/zenodo.6053952
3. Given the publication dates, we are limited in our analysis to the poetry of *early* modernism.
4. Hyperparameter: 500 steps, batchsize 30, learningrate 2e-5 (see Konle and Jannidis 2020, Gururangan et al. 2020)
5. Very frequent emotions like longing (f1: 0.73) or suffering (f1: 0.72) yield sufficient classifiers, but less frequent ones like calmness or desire lead to results similar to a random baseline.
6. In our study, in accordance with German literary histories, Stefan George (22 poems), Rainer Maria Rilke (37 poems), Hugo von Hofmannsthal (31 poems), and Arno Holz (50 poems) represent canonical modernism.

Accuracy is not all you need

Kristensen-McLachlan, Ross Deans
rdkm@cas.au.dk
Aarhus University, Denmark

Lassen, Ida Marie S.
idamarie@cas.au.dk
Aarhus University, Denmark

Enevoldsen, Kenneth
kenneth.enevoldsen@cas.au.dk
Aarhus University, Denmark

Hansen, Lasse
lasse.hansen@clin.au.dk
Aarhus University, Denmark

Nielbo, Kristoffer L.
kln@cas.au.dk
Aarhus University, Denmark

Introduction

Discussions around diversity and bias in language representations are a hot topic in contemporary natural language processing. Countless papers have pointed out that these representations can be shown to contain specific biases, such as in the case of both so-called static

embeddings (Bolukbasi et al., 2016) and more state-of-the-art contextual approaches (Zhao et al., 2019).

In this paper, we contribute to on-going attempts to quantify and measure the effects of this bias for specific NLP tasks. We present results of an experiment which tests the effect of data biases on the performance of different NLP frameworks for named entity recognition (NER) in Danish. While the immediate results deal specifically with only this language, the methods employed can be adapted to other linguistic and cultural contexts, given appropriate modifications. We choose to focus specifically on biases in the sense of representational harms (Barocas et al., 2017) by investigating the performance differences of NER in Danish for different social groups, namely gender and ethnicity (Shah et al., 2019). Despite this quite narrow focus, we aim to contribute to wider discussions in the field of *fairML* and bias in NLP.

Recent work has pointed out how unintended bias in NLP systems leads to systematic differences in performance for different demographic groups (Borkan et al, 2019). Building on these insights, frameworks, fairness metrics and recommendations for the field have been developed to quantify and mitigate bias (Borkan et al., 2019; Shah et al., 2019; Zhao et al., 2019; Blodgett et al., 2020; Czarnowska et al., 2021). More specifically relevant for this work, Zhao et al. (2018) have shown how data augmentation on training data can eliminate gender bias for coreference resolution. With our work, we propose another use of data augmentation, namely as a method to test the robustness of NLP models as well as uncover potential social biases in the model.

Method

We define bias as systematic difference in error – *error disparity* – as a function of a given sensitive features (Shah et al. 2019). In other words, bias in the model is measured through difference of performance accuracy when data is augmented with different gender and ethnicity features.

In Enevoldsen et al. (2021) a range of contemporary Danish NLP frameworks were subjected to a series of data augmentation strategies to test their robustness. These augmentations included random keystroke augmentation to simulate spelling errors; and spelling variations specific to the Danish language. Additionally, among the augmentation strategies were the following name augmentations:

1. Substitute all names (PER entities) with randomly sampled Danish names, respecting first and last names.
2. Substitute all names with randomly sampled names of Muslim origin used in Denmark (Meldgaard, 2005), respecting first and last names.
3. Substitute all names with sampled Danish male names, respecting first and last names.
4. Substitute all names with sampled Danish female names, respecting first and last names.

The purpose of these augmentations was to test specifically the robustness of named entity recognition in each of the Danish NLP frameworks given data which had been augmented relative to gender and ethnicity. If a framework performed just as well (or better) with these augmentation as without, this is taken to indicate robustness. If a framework performs worse, we are able to quantify exactly where the model is failing and, hence, where potential biases reside.

Table 1 illustrates general performance of contemporary Danish NLP frameworks on a range of tasks. We can see that larger, transformer-based models consistently outperform other models, particularly on NER tasks. At a general level, these results underline three well-known trends in deep learning and NLP: 1) larger models tend to perform better; 2) higher quality pre-training data leads to better models; and 3) multilingual models perform competitively with monolingual models (Brown et al., 2020; Raffel et al., 2020; Xue et al., 2021).

Results

The results from the full range of augmentation strategies can be found in Enevoldsen et al. (2021). In Table 2, we see only the results following name augmentation. From this table, we can see that the NER performance of every model is affected to some degree by the data augmentations. What is immediately apparent, though, is that not all models are affected equally and not all augmentations cause as pronounced effects. Our results seem to demonstrate that Danish language models are relatively robust to the effects of gender. However, the same can not be said for Muslim names, which cause significantly worse performance for all models. This suggests that Danish NLP models contain a greater relative bias in terms of ethnicity than gender.

Table 1
General performance of Danish NLP frameworks. Wall Time is the time taken by the model to go through the DaNE test set without augmentation. Empty cells indicates that the framework does not include the specific model.

	POS	NER						Dependency Parsing		Wall Time
Model	Accuracy	Person	Location	Organization	Misc	F1	F1 w/o Misc	LAS	UAS	GPU/CPU
DaCy large	**98.37**	**93.33**	**84.88**	**76.49**	**80.16**	**84.39**	**85.65**	**88.44**	**90.85**	2.9 / 34.7
DaCy medium	98.15	89.86	83.96	64.47	70.09	77.67	79.68	86.65	89.25	1.8 / 9.9
DaCy small	97.75	87.98	79.23	60.58	64.82	74.18	76.98	84.03	87.63	1.9 / 2.6
DaNLP BERT		92.27	83.90	71.13		72.84	83.20			37.4 / -
Flair	97.80	92.60	84.82	61.29		70.49	81.09			2.0 / -
NERDA		92.35	81.52	65.96	72.41	79.04	80.85			2.5 / -
Polyglot	76.26	79.25	68.06	40.69		56.67	65.32			- / 3.8
SpaCy large	96.30	86.17	84.16	63.36	65.52	75.75	78.57	78.01	81.95	0.9 / 1.4
SpaCy medium	95.71	84.55	77.29	63.16	63.25	73.23	76.01	77.73	81.87	1.2 / 1.4
SpaCy small	94.80	78.92	69.04	53.49	61.54	67.11	68.61	74.03	78.68	1.4 / 1.5
Stanza	97.62							83.84	87.34	29.3 / -

Table 2
Named Entity Recognition (NER) performance of Danish Natural Language Processing (NLP) pipelines reported as average F1 scores excluding the MISC category on the test set. Best scores are marked bold and second best are underlined. * denotes that the result is significantly different from the baseline using a significance threshold of 0.05 with Bonferroni correction for multiple comparisons. Danish names is considered the baseline for the augmentation of Muslim, female, and male names. Values in parentheses denote the standard deviation. NERDA limits input size to 128 wordpieces which leads to truncation on long input sizes and high rates of keystroke errors.

	Names			
Model	Danish	Muslim	Female	Male
DaCy large	**86.2 (0.6)***	**86.0 (0.5)**	**86.2 (0.5)**	**86.2 (0.4)**
DaCy medium	80.3 (0.5)*	77.9 (0.8)*	80.3 (0.4)	80.2 (0.7)
DaCy small	76.5 (0.9)	75.7 (0.7)*	76.7 (0.8)	76.6 (0.7)
DaNLP BERT	82.9 (0.6)	81.0 (1.0)*	83.1 (0.5)	83.0 (0.7)
Flair	81.2 (0.7)	79.8 (0.7)*	81.4 (0.5)	81.5 (0.5)
NERDA	80.0 (1.1)*	78.1 (1.2)*	80.2 (0.8)	80.0 (0.8)
Polyglot	63.1 (1.2)*	41.8 (0.7)*	61.2 (1.2)*	64.8 (1.2)*
SpaCy large	79.5 (0.6)*	71.6 (1.1)*	79.8 (0.5)	79.4 (0.5)
SpaCy medium	78.2 (0.7)*	69.2 (1.4)*	78.2 (0.7)	78.5 (0.8)
SpaCy small	62.5 (1.6)*	57.8 (1.4)*	63.0 (1.1)	63.3 (0.9)

Conclusions

There are some notable limitations to the current study. Firstly, we have presented experimental results for a single, comparatively small Indo-European language. Nevertheless, we predict that similar results could be obtained for different languages, given an appropriate change of experimental conditions. Secondly, by making augmentation on male and female names we continue the folk conception of gender, where gender is understood as binary and static. Finally, we have not further separated Muslim names into typically male Muslim names and typically female Muslim names. Were our results stratified in this manner, we would be in the position to conduct a more intersectional analysis into the relative effect of gender and ethnicity. This more nuanced perspective is the goal of a future experiment.

However, the present results already paint a complex but essentially coherent picture. We have demonstrated that data augmentation is a simple and transparent way of testing the robustness of contemporary language models on tasks like named entity recognition. Moreover, we have shown that these data augmentation tasks can be used to test for specific biases in language models with respect to categories such as gender, race, and ethnicity. Doing so puts us in the position to evaluate the performance of specific language models not only in relation their overall performance metrics such as accuracy and micro-F1 scores but also their robustness and biases relative to these categories.

Our proposal is that in the domain of NLP and language modelling *accuracy is not all you need.* This is especially true for researchers who apply NLP tools for text analysis in the humanities and social sciences. For example, a researcher working in the field of gender history might need their models to be particularly robust with respect to gender; a scholar of social media might have a specific reason to require that their model is particularly robust to different ethnicities represented in their data. Rather like the trade-off between precision and recall or between speed and accuracy, there may need to be a trade-off between the overall accuracy of a particular language model and its robustness in terms of diversity. By reflecting on bias in model selection we advocate for a move of responsibility for model biases back to the technologists developing and deploying these models instead of pushing it to the communities affected by biases in NLP systems (Blodgett et al., 2020).

Bibliography

Barocas, S., Crawford, K., Shapiro, A., & Wallach, H. (2017). The problem with bias: Allocative versus representational harms in machine learning, Proceedings of SIGCIS 2017.

Blodgett, S. L., Barocas, S., H. D. III, & Wallach, H. M. (2020). Language (technology) is power: A critical survey of "bias" in NLP, URL: https://arxiv.org/abs/2005.14050. arXiv:2005.14050.

Bolukbasi, T., Chang, K., Zou, J. Y., Saligrama, V. & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings, URL:http://arxiv.org/abs/1607.06520. arXiv:1607.06520.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., & Vasserman, L. (2019). Nuanced metrics for measuring unintended bias with real data for text classification, URL: http://arxiv.org/abs/1903.04561. arXiv:1903.04561.

Brown, T. B. , Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R, Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020) Language models are few-shot learners. URL: http://arxiv.org/abs/2005.14165. arXiv:2005.14165.

Czarnowska, P., Vyas, Y., & Shah, K. (2021). Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics, URL: https://arxiv.org/abs/2106.14574. arXiv:2106.14574.

Enevoldsen, K., Hansen, L., & Nielbo, K. (2021). Dacy: A unified framework for Danish NLP, URL: https://doi.org/10.48550/arXiv.2107.05295, arXiv:2107.05295.

Gaut, A., Sun, T., Tang, S., Huang, Y., Qian, J., ElSherief, M., Zhao, J., Mirza, D., Belding, E. M., Chang, K., & Wang, W. Y. (2019). Towards understanding gender bias in relation extraction, URL: http://arxiv.org/abs/1911.03642. arXiv:1911.03642.

Meldgaard, E. V. (2005). Muslimske fornavne i danmark. URL: https://nors.ku.dk/publikationer/webpublikationer/muslimske fornavne/, Københavns Universitet.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020) Exploring the limits of transfer learning with a unified text-

to-text transformer, URL: http://arxiv.org/abs/1910.10683. arXiv:1910.10683.

Shah, D., Schwartz, H. A., & Hovy, D. (2019). Predictive biases in natural language processing models: A conceptual framework and overview, URL: http://arxiv.org/abs/1912.11078. arXiv:1912.11078.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2020) mT5: A massively multilingual pre-trained text-to-text transformer. URL: http://arxiv.org/abs/2010.11934. arXiv:2010.11934.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods, URL: http://arxiv.org/abs/1804.06876. arXiv:1804.06876.

Zhao, J, Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., & Chang, K.W. (2019). Gender bias in contextualized word embeddings, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, pp. 629–634. URL https://aclanthology.org/N19-1064. doi:10.18653/v1/N19-1064.

Semantically-Grounded Generative Modeling of Chinese Landscapes

Kulyabin, Mikhail
mikhail.kulyabin@fau.de
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Germany

Kosti, Ronak
ronak.kosti@fau.de
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Germany

Bell, Peter
peter.bell@uni-marburg.de
Philipps-Universität Marburg, Germany

Motivation – Chinese landscape paintings have unique and different visual features and stylistic characteristics as compared to the western counterpart. At the thematic level, they depict the local traditions, culture, artistic movement and the geographic variety offered by the Asian subcontinent. And at the image level, the paintings depict low texture variations due to the particular handling of the brush-stroke, color, tone and the artistic genre. These landscapes, therefore, provide valuable insights into the structural and stylistic aspects of the paintings. The state of the art methods in image analysis enables finding semantic correlations for analysis across huge image collections, and also often constitute iconographical analysis. For example, the digitization of Dunhuang grottoes (a world heritage site) has led to immense knowledge creation in terms of iconographical art-historical understanding of Panofsky (Wang et al., 2018) for the categorical and semantic treatment of the objects present in the grottoes, as well as in the computational community for discovering the ruined parts of the murals via automatic restoration techniques. We aim to investigate the semantic aspect of the Chinese landscape paintings computationally using conditional control of the image generation.

Learning basic structure of the painting and the use of artistic style helps in getting conditional control over the generation of similar looking paintings. This allows us to get a closer reading of the formation process behind those landscapes. Generative adversarial networks (GAN) (Gatys et al., 2016) are a type of deep neural networks - when given a collection of images, they are able to learn the styles from them and are able to generate images with similar styles. For example, a GAN trained on a set of paintings with Impressionism style, will be able to generate images with the same style.

Related Work – Generative modeling has thus been applied to various artworks for a variety of tasks like artwork synthesis, image editing, style transfer (Gatys et al., 2016), and image-to-image translation. Previous work on generating landscapes mainly focused on image-to-image translation, going from sketches to landscape generation. These techniques use input conditions like ink wash tone, brush strokes or a sketch as conditional input for artwork generation. However, these popular neural style transfer methods do not work well with Chinese artworks since there are marked differences in the depiction of textures, abstraction, structure and style of the paintings. Another way of controlling the generative aspect is to use input semantic maps (Liu et al., 2019) which accentuate different foreground and background objects as conditional inputs.

Research Gap – The main problem with the above methods is that the generative networks are often caught in degenerate solutions which makes the networks generate images with limited variety. Although training with image-to-image translation networks uses the image itself or its sketch as a conditional input in order to capture more diverse aspects of the paintings, the generated images often render finer details of various foreground objects unclear. When only sketches are used, the GANs are not able to capture the color-texture distribution of the generated objects. They tend to mix or use colors that are not consistent with the true colors. Semantic maps of each input image can help to bring color-consistency in image

generation and also provide control to generating specific objects or regions of interest in the final image. However, these semantic maps are difficult to obtain since they require manual annotation.

Proposed Approach – In our work, we demonstrate generative modeling of Chinese landscapes using sketches as well as semantic segmentation maps of the corresponding input paintings. Specifically, we propose a novel way of generating semantic segmentation maps without requiring any kind of manual annotations. The sketches and segmentation maps of the paintings are not easily available, therefore obtaining a good quality segmentation map is a big challenge. Our approach is divided into three stages:

In the first stage, we generate sketches of Chinese landscape paintings using an edge detector called HED. This edge detector is chosen since it generates high-level shapes of the image structures while also retaining low-level details. Next, we use watershed for image segmentation to generate weak segmentations. The algorithm does not work equally well for all images due to the variety of abstractions in shapes and styles in the landscape images.

In the second stage, we refine the weak segmentations by first color-coding them. The color-codes are chosen based on the average color of the corresponding overlapping region with the input landscape image. These maps are subdued in their color tones, so we color equalize to increase their contrast. Then, these maps are taken together with the corresponding sketches to train a U-Net (similar to the generator of Pix2Pix GAN) for multi-task optimization. The landscape paintings are used as inputs and the segmentation maps along with the sketches are the outputs, while optimization conditions are constrained to only sketches. Optimizing for sketches helps to improve the quality of segmentation maps based on the structure of the sketches. Subsequently the network is able to generate better segmentation with color-consistency.

In the third and final stage, we use the Pix2Pix GAN to generate similar samples of Chinese landscape paintings from input sketches and segmentation maps. A previous work (Xue et al., 2021) explores a similar method, but uses only sketches. We use high-quality segmentation maps in addition for better foreground region generation. The training time is huge and depends mainly on the dataset size and the input image size. Higher input image size normally tends to increase the quality of the finer objects, but also takes longer training times.

Significance of Our Approach - We show how to generate better quality semantic segmentation maps using a HED detector, watershed algorithm and an additional training stage with U-Net to refine the weak segmentations generated by the watershed algorithm. In all of these steps, there is no requirement of additional high-quality manual labels. Therefore, our approach can be applied to any new set of images for generating better segmentation maps.

Bibliography

Gatys, L. A., Ecker, A. S. and Bethge, M. (2016). Image Style Transfer Using Convolutional Neural Networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, pp. 2414–23 doi: 10.1109/CVPR.2016.265.

Liu, X., Yin, G., Shao, J. and Wang, X. (2019). Learning to predict layout-to-image conditional convolutions for semantic image synthesis. *Advances in Neural Information Processing Systems*, *32*.

Wang, X., Song, N., Zhang, L. and Jiang, Y. (2018). Understanding subjects contained in Dunhuang mural images for deep semantic annotation. *Journal of Documentation*, **74**(2): 333–53 doi: 10.1108/JD-03-2017-0033.

Xue, A. (2021). End-to-End Chinese Landscape Painting Creation Using Generative Adversarial Networks. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Waikoloa, HI, USA: IEEE, pp. 3862–70 doi: 10.1109/WACV48630.2021.00391.

From Roland to Conan: First results on the corpus of French literary fictions (1050-1920)

Langlais, Pierre-Carl
pierrecarl.langlais@gmail.com
Université de Montpellier Paul-Valéry

Camps, Jean-Baptiste
jean-baptiste.camps@chartes.psl.eu
École nationale des chartes | Université PSL, France

Baumard, Nicolas
nbaumard@gmail.com
École normale supérieure, PSL

Morin, Olivier
alf.drummond@gmail.com
MPI für Menschheitsgechichte, Jena

A corpus of French literary fictions: 1050-1920

A decade ago, when he set out to explore the evolution of English novels in the period 1740–1850, F. Moretti had to focus on the titles as the main available data, yet noting that:

> in a few years, we will have a digital archive with the full texts of (almost) all novels ever published .

Today, we propose to embark on a journey to model the evolution of French literary fictions, from their epic and chivalric origins up to the more modern productions of genres such as heroïc fantasy or historical novels. We offer to do that not from the titles, but by analysing a corpus, currently in construction, whose aim is to cover all French literary fiction digitized from the earliest sample of French literature to the 20th century (or more precisely, to the point were a significant share of published literature is not yet in the public domain).

This project is at the intersection of two major trends in computational literary analysis: the creation and documentation of large literary corpora and the analysis of literary genre and discourse through machine learning classification. In comparison with previous examples of French literary corpus (like *théâtre classique*) or the French corpus of the European literary text collection , French novels make up a massive amount of texts (80,000 registered work in the French National Library before the 20th century), a large share of which is non-canonical and little documented. Text mining techniques make it possible to explore and document large digitized corpora with little editorial work. Classification is not simply used as cataloguing tool: its limitations can in fact inform in a more complex way the development of genres and the intertextual interplay between one genre and another.

The French fiction corpus initially results from the collocation of three different collections:

1. A collection of medieval fictions and *chansons de gestes* (1050-1450) - (cf. Camps et al., 2019).
2. A collection of printed fictions of Gallica from the modern period and the 19th century (Langlais, 2021b).
3. A new collection comprising most of the digitized fictions from the early modern period (1450-1700).

The second corpus relied on one of the oldest bibliographic database: the catalog of the French National Library. The classification scheme used by the library until 1996 was developed between 1684 and 1688 by the librarian of Louis XIV, Nicolas Clément. A specific category for novels (Y2) has been existing since 1730 as a duplication of the category for poetry (Y). The novels are now classified in Y2 or Y Bis. For cultural history, the BNF catalog presents a major interest: the categories are often (nearly) contemporary of the documents they aim to classify.

This corpus was made possible thanks to the policy of open data and open content which the BNF has been engaged in for several years. While the Clément classification ceased to be used for the classification of physical collections in 1996, it has become available in 2017 for data analysis when the BNF opened a new catalog access service, the SRU [1]. It was originally limited to novels digitized by the French National Library with a strong focus on the 19th century.

The third corpus was created specifically for the project: it aims to cover all the digitized versions of early modern novels published in France between 1473 (publication date of our earliest entry, *Le Roman de Jason et Médée*) and 1700 (our arbitrary cutoff, for now). This collection originally aimed to bridge the two previous corpora which were focused on two extreme temporal points of the history of French literature (the medieval period and the post-1800 period). It became a more ambitious experiment: a systematic harvesting of all available digitizations. The collections of digital documents already available in *Fictions littéraires de Gallica* have been significantly expanded thanks to the semi-automatical retrieval of documents digitized by Google Books and other online digital library. The creation of this composite corpus underline that digitized collections may be already more representative than expected, although nobody knew it or could measure the scale of it.

Using the combined collections of several digital libraries made it possible to cover a large amount of the novels registered in the catalog of the French National Library. For the 1470-1600 period, we have retrieved 275 novels out 349, that is 78.8% of the corpus. For the 1600-1700 period, we have 724 novels out of 1058, that is 68.4% of the corpus. Representativeness ratio is not only high but consistent on the entire time period as shown in figure as the size of the available digitized corpus remains proportional to the total amount of novels identified by the French National Library.

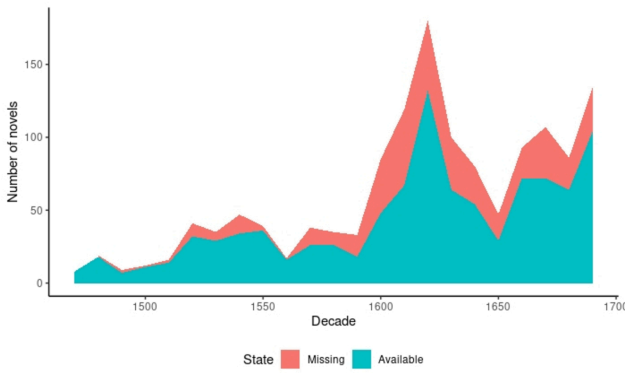


Figure 1:
Composition of the corpus by decade

Such high coverage seems to alleviate most concerns related to the representativeness of digital corpora: it becomes less likely that important genres or themas are neglected. Yet the novels registered on the French National Library catalog do not encompass the total sum of literary fiction published or circulated on the period. The comparison with Google Books showed that numerous editions were not recorded by Gallica. In the context of the project, we only checked novels recorded on the catalogue, but this discrepancy shows that some published novels could definitely have been overlooked, although by definition these documents are not necessarily expected to belong to those with the largest impact, as they failed to be noticed by bibliographic records. After the 19th century, nearly all published monographs are expected to be indexed in the catalogue of the French National Library due to progress in the implementation of the policy of legal deposit (*dépôt légal*, established in 1537). Yet, at this time, a large amount of literature was increasingly being published in periodicals and newspapers.

In short as representativeness of existing library catalogues becomes less of a concern, digital collections are bound to raise more complex issues and address a more critical perspective on existing bibliographic resources: what is a "novel"? what is a "publication"? what about piracy edition or periodical fictions?

The creation of the third corpus also aims to benefit from unprecedented progresses in the detection of historical OCR. The OCR17+ model trained by Claire Jahan and Simon Gabay (2021) on a collection of 17th century prints already yields a usable text for most of the 16th to 18th century documents [2].

Modelling literary genre: a back and forth approach

Chivalric romance is one of oldest and most enduring genre of European literature. It finds its origin in Old French epics, known as *chansons de geste*, whose first attestations go back to the 11th century, and in the genre of *roman* that emerges in the second half of the 12the century. Continuously modified, adapted and rewritten during the Middle Ages and the Early Modern times, this repertoire of fictions provided numerous tropes and patterns that arguably live on in more modern and even contemporary literature, especially in popular literature and even fantasy novels.

Preliminary exploration of our large corpus of French literary fictions suggests that the chivalric novel, rather than ceasing to exist, has been continually evolving and gradually morphed into forms close to the contemporary fantasy novels. This structural transformation has been largely overlooked as forms of chilvalric romances have largely disappeared from the high literary canon after the Renaissance and continued to evolve in the production of lesser known and more obscure authors.

Our analysis relies on a anachronistic use of classification recently pioneered in cultural analytics . Probability rates, cross-classifications of the same text and, even, classification failures are reinterpreted as a way to measure the complex evolution of literary genres.

To investigate the transformation of chivalric novels into contemporary genres like fantasy, we created two historical models of literary genres: a 21st century model and an early modern "Fresnoy" model (from a 1731 catalog). The combined use of anachronistic classification aims to locate the two parallel processes of genre survival/transformation (for chivalric romance) and genre emergence/coagulation (for fantasy).

Classification was made with SVM using an R library initially developed for the classification of newspaper genres in the 19th century, *TidySupervise* .

The 21st century model is created using nine genre categories of the social cataloguing website *Babelio* [3]. User-generated tags have recently emerged as an important source of information in computational analysis of contemporary literature and literary reception . A French counterpart to *Goodreads*, *Babelio* has a significant impact among French literary readers with nearly 1 million visits per month. In this project, we focused on a subset of generic tags that were either major acknowledged genres in contemporary literature or relevant for our ongoing projects: romance, fantasy, detective fiction, science fiction, historical novel, adventure novel, social novel, *fantastique* novel and erotica. Obviously, this is not a straightforward classification, since one novel could belong to several categories. We aimed

rather to reconstruct the fuzzy space of contemporary literary genres in France with all its underlying uncertainties and overlaps.

The model was trained on 4,081 segments of 1000 words extracted from 1,346 novels. We applied a random selection of three 1000 words segments by work. This selection aims to limit over-fitting and ensure that the model will be correctly trained on generic features and not on the style of specific novels. Bootstrap evaluation of the model yield a 75% accuracy, yet with significant variations among the genres (fantasy being the highest rated with science fiction and erotica).

Preliminary results from the 21st century model have revealed significant examples of "missing links" between the early modern chivalric romances and the fantasy novels (Table 1). Of special interest is *La Mort de Roland* a 1858 rewrite of the *Chanson de Roland* by Alfred Assolant that explicitly claims to be an *epic fantasy* (*Fantaisie épique*).

Title	Author	Segments classified	Proportion of segments
Les rois de mer = *The Sea Kings*	Léon Cahun	74	70
Iskender	Judith Gauthier	58	66
Les aventures du dernier Abencérage = *The Adventures of the last Abencérage*	François-René de Chateaubriand	15	66
Les aventures du capitaine Magon = *The Adventures of capitain Magon*	Léon Cahun	108	65
Histoire de Don Quichotte racontée à la jeunesse = *The Story of Don Quichotte told to the Youth*	Miguel de Cervantes	55	65
Voyage de Mademoiselle Lili autour du monde = *The Travel of Miss Lili around the world*	P.-J. Stahl	14	64
Le petit duc = *The Small Duke*	Charlotte Mary Yonge	43	63
Guillaume Tell [de Schiller], adapté pour les enfants = *Schiller's William Tell, adapted for Childrens*	Friedrich von Schuller	17	58
La mort de Roland = *The Death of Roland*	Alfred Assolant	70	58
La flèche noire = *The Black Arrow*	Robert Louis Stevenson	81	55

Table 1.
Top 10 works classified as fantasy in the corpus of 19th century novel digitized by Gallica

Most of these works are poorly attested in literary history. In comparison, the results from Science-fiction yields much more expected and "canonic" works (especially from Jules Verne).

The early modern model (or "Fresnoy" model) has been made possible by an exceptional historical source on literary genre classification: the second volumes of *De l'usage des romans, où l'on fait voir leur utilité & leurs différens caractères* by Nicolas Lenglet du Fresnoy (first published in 1731 under the pseudonym of Gordon de Percel). It is a catalog of a large among of French, Spanish and Italian novels published since 1731 broken down by genres according to the prevalent taxonomy of the time: *Roman de chevalerie*, *Roman d'amour*, *Roman historique*, *Roman comique*, *Roman politique*, etc. Ongoing work aims to reconcile the classification of Lenglet du Fresnoy with our corpus of digitized novels. Preliminary results suggest that the generic identity in the catalogue of Fresnoy is much stronger than in the Babelio dataset, with as much as 93% accuracy on four genres (*Roman d'amour*, *Roman historique*, *Roman de chevalerie* and *Roman comique & satirique*) in our initial run. While this high accuracy may be caused by overfitting on a limited samples of novels, it seems also consistent with the significance of genre classification in the meta-discourse about the novel in the 18th century.

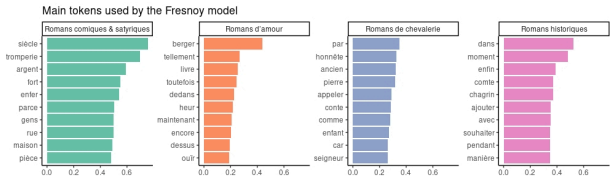


Figure 2:
Main words used in the four categories of the Fresnoy model

A this stage, the expansion of the corpus and the re-digitization of available documents with historical models for OCR will be crucial to move beyond an exploratory phase and design a more systematic examination of the metamorphosis of literary genre.

Bibliography

Camps, J.B. *et al.* **(2019).** *Geste: un corpus de chansons de geste, 2016-... (Version 02)*. Paris. URL http://doi.org/10.5281/zenodo.2630574.

Fièvre, P. (2007). Théâtre classique. URL http://www.theatre-classique.fr/.

Jahan, C. and Gabay, S. (2021). *OCR17+ - Layout analysis and text recognition for 17th c. French prints*, Paris/Genève: ENS Paris/UniGE, https://github.com/e-ditiones/OCR17plus.

Langlais, P.C. (2021). Classified News, Redefining the history of newspaper genre with supervised models. In *Digital Newspaper: a new Eldorado for the historians*. De Gruyter.

Langlais, P.C. (2021 b). *Fictions littéraires de Gallica / Literary fictions of Gallica*, URL https://doi.org/10.5281/zenodo.4751204.

Moretti, F. (2009). Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850), *Critical Inquiry*, 36(1):134–158, 2009. doi: 10.1086/606125. URL http://www.jstor.org/stable/10.1086/606125.

Odebrecht, C. , Burnard, L. , and Schöch, C. (2021). European literary text collection (eltec).

Calvo Tello, J. (2021). *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning*. Bielefeld University Press. doi: 10.1515/9783839459256.

Underwood, T. (2019). *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press.

Walsh, M. and Antoniak, M. (2021). The goodreads “Classics”: A computational study of readers, amazon, and crowdsourced amateur criticism. *Journal of Cultural Analytics*, 6(2). doi: 10.22148/001c.22221.

Notes

1. *Search/Retrieve via URL*, https://api.bnf.fr/fr/api-sru-catalogue-general.
2. Jahan and Gabay (2021). This model will be applied on the entire corpus. Preliminary results on a random selection of 60 pages from the corpus show that OCR quality is already high for the 17th century (nearly 99% character accuracy). The 16th century corpus is more challenging due to specific issues with Gothic fonts and segmentation, although we intend to solve these limitations by fine tuning a model for Gothic fonts in incunabula and 16th century prints. In comparison, the mean OCR quality in the collections held by Gallica is at roughly 80% for the 16th and 17th centuries (according to available metadata) and a large part is currently unavailable for full-text search, possibly because OCR quality was too low. The project will also benefit from the development made by the ongoing Gallic(orpor)a project, that aims to provide a fully reusable pipeline for Gallica documents in French from the Middle Ages to the Revolution. By the end of this project, we plan to create a small search engines of the entire collected corpus of 15th to 17th century documents. This would be the first resource to give access to a large share of the published literature in France in the early modern period. An initial version of this search engine may be available by the time of the conference and serve as a demonstration of our methodology of systematic collection and enhancement of available digitized collections.
3. https://www.babelio.com/.

Newspaper Navigator: Reimagining Digitized Newspapers with Machine Learning

Lee, Benjamin Charles Germain
bcgl@cs.washington.edu
University of Washington, United States of America

The millions of digitized historic newspaper pages within *Chronicling America*, a joint initiative between the Library of Congress and the National Endowment for the Humanities, represent an incredibly rich resource. Historians, journalists, genealogists, students, and members of the public explore the collection regularly via keyword search. But how do we navigate the abundant visual content in *Chronicling America*? This question is motivated by the fact that visual culture within newspapers has proven to be a capacious source for humanists. Within periodicals studies, scholars have utilized the visual content in newspapers to investigate topics as far ranging as the evolution of comedic sensibilities within comic strips to hidden editorial practices embedded within newspaper layout (Cole, 2020; Barnhurst and Nerone, 2002). ‚ This collective body of work is bolstered by new methodologies being employed within the digital humanities to extract and analyze visual content in historic newspapers (Piper, Wellmon, and Cheriet, 2020; Fyfe and Ge, 2018; Wevers and Smits, 2020). „ In this talk, I will present my project, *Newspaper Navigator*, created in collaboration with LC Labs, the National Digital Newspaper Program, and IT Design & Development at the Library of Congress, as well as Professor Daniel Weld at the University of Washington. In particular, I will discuss four distinct phases of *Newspaper Navigator* to extract and analyze the visual content within *Chronicling America* and beyond.

First, I will describe extracting visual content, including photographs, illustrations, comics, editorial cartoons, maps, headlines, and advertisements, from 16.3 million pages in *Chronicling America*, resulting in the *Newspaper Navigator* dataset . To accomplish this, I finetuned an object detection model of thousands of bounding box annotations of visual content from the *Beyond Words* crowdsourcing initiative launched by LC Labs in 2017. I then made a full pass over 100TB of image and XML data in order to construct the dataset. The Library of Congress and I released the resulting *Newspaper Navigator* dataset to the American public in May, 2020, as the largest dataset of its kind ever produced. In pursuit of the Library’s mission of improving access, we placed the dataset and all code into the public domain for unrestricted re-use. We published a paper describing the dataset and its construction at the 2020 ACM Conference on Information Knowledge & Management (Lee et al., 2020).

Second, I will discuss the *Newspaper Navigator* public search application for 1.5 million photos from the dataset. While caption-based keyword search for images provides much utility, the approach also has fundamental limitations: for example, how do historians search for photographs with distinct visual motifs? This question is particularly relevant for cultural heritage collections, where OCR transcriptions are inevitably imperfect, further restricting the efficacy of keyword search. In the second phase of *Newspaper Navigator*, I created and deployed the search application

for 1.5 million photographs in the dataset based on the real needs that historians and other users had articulated to us surrounding these limitations. In addition to providing keyword search functionality, the search application enables users to iteratively train machine learning algorithms in order to retrieve visually similar photos according to topics or concepts of interest, such as baseball players. From an exploratory search perspective, I call this search functionality *open faceted search* because it empowers users to create their own facets dynamically, facilitated by interactive machine learning algorithms that can train and predict over all 1.5 million photos in under a second. Unlike standard faceted search, open faceted search provides a path forward even when metadata is impoverished, making it extensible to a wide range of digitized collections. I first presented open faceted search in a demo at the 2020 ACM Symposium on User Interface and Software Technology (Lee and Weld, 2020).

Third, I will discuss the *Newspaper Navigator* data archaeology, which I wrote to examine the ways in which a *Chronicling America* newspaper page is altered and decontextualized during its journey from a physical artifact to a series of probabilistic photographs in *Newspaper Navigator*. First released with the *Newspaper Navigator* search application in order to provide scholars and the general public alike with a resource surrounding the ethical considerations and implications of this project, the data archaeology has appeared in revised form as an article in *Digital Humanities Quarterly* (Lee, 2021). In this data archaeology, I studied the digitization journeys of four different pages in Black newspapers in *Chronicling America* that reproduce the same photograph of W.E.B. Du Bois. In tracing the pages' journeys, I unpacked how each step, from microfilming to OCR to image embeddings, propagates bias, marginalization, and erasure via the machine learning algorithms employed.

I will conclude by discussing *Newspaper Navigator* research collaborations with scholars and educators across universities and cultural heritage institutions. With Devin Naar, I conducted the first study of the Ladino press at a macroscopic scale. Ladino, also known as Judeo-Spanish, is the language of the Sephardic Jewish people, and the Ladino press represents an invaluable source for studying Sephardic Jewish experiences across the world. In this collaboration, I have utilized *Newspaper Navigator* to excavate the visual content from over 15,000 pages of Ladino newspapers. Many Ladino texts are not even keyword searchable due to the widespread failure of OCR engines to properly transcribe the language. My excavation of the visual content offers the first path forward to studying Ladino newspapers at scale and thus serves as a corrective to this algorithmic marginalization. My analysis of thousands of extracted photographs and advertisements reveals new contours to Sephardic Jewish experiences in modernity: in addition to uncovering photographs of individuals and communities, I have also identified an abundance of advertisements offering remedies for anxieties, whether medical, financial, or class-based. My findings are detailed in my chapter in *Jewish Studies in the Digital Age*, currently in press with De Gruyter Press for publication in 2022 (Lee, 2022).

Moreover, in an ongoing collaboration with periodicals scholars Jim Casey, Sarah Salter, and Joshua Ortiz Baco, I am studying the evolution of visual layouts of newspaper titles, with a particular focus on ethnic presses and how they served as vehicles for protest and community. Using the *Newspaper Navigator* dataset, it is possible to directly quantify the similarity of layouts across millions of newspaper pages, enabling us not only to trace the technological developments of printing presses but also to uncover the hidden editorial practices embedded within layouts themselves. For example, we have identified clusters of newspaper titles with similar visual layouts, such as networks of African-American titles that feature illustrations and photographs of members of their communities in portrait poses in the center of their front pages. The editors' choice of a shared visual grammar speaks to the ways in which visual culture and layout featured prominently into editorial practices. We presented our first paper detailing this collaboration at the *Computational Humanities Research* 2021 conference (Lee et al., 2021). Lastly, I have collaborated with professors of education Ilene Berson and Michael Berson to investigate uses of *Newspaper Navigator* in the classroom, as detailed in our article in *Social Education* (Lee, Berson, and Berson, 2021).

I will conclude my talk by reflecting on possibilities for research at the intersection of machine learning, the digital humanities, and libraries.

Newspaper Navigator **Resources:**

- *Newspaper Navigator* dataset:https://news-navigator.labs.loc.gov/
- *Newspaper Navigator* search application: https://news-navigator.labs.loc.gov/search
- *Newspaper Navigator* data archaeology: https://hcommons.org/deposits/item/hc:32415
- *Newspaper Navigator* project description & other links: https://bcglee.github.io/newspaper-navigator.html

Bibliography

Barnhurst, K. G. and Nerone, J. (2002). *The Form of News: A History*. Guilford Press.

Cole, J. L. (2020). How the Other Half Laughs: The Comic Sensibility in American Culture, 1895-1920. University Press of Mississippi.

Fyfe, P. and Ge, Q. (2018). Image Analytics and the Nineteenth-Century Illustrated Newspaper. *Journal of Cultural Analytics*, **3**(1). DOI: 10.22148/16.026.

Lee, B. C. G. (2021). Compounded Mediation: A Data Archaeology of the Newspaper Navigator Dataset. *Digital Humanities Quarterly*, **15**(4). DOI: http://www.digitalhumanities.org/dhq/vol/15/4/000578/000578.html.

Lee, B. C. G. (2022). The Digital Humanities and the Ladino Press: Using Machine Learning to Extract and Analyze Visual Content in Historic Ladino Newspapers. *Jewish Studies in the Digital Age*. De Gruyter Press.

Lee, B. C. G., Baco, J. O., Salter, S. H. and Casey, J. (2021). Navigating the Mise-en-Page: Interpretive Machine Learning Approaches to the Visual Layouts of Multi-Ethnic Periodicals. *Computational Humanities Research Conference* 2021. DOI: http://arxiv.org/abs/2109.01732.

Lee, B. C. G., Berson, I. R. and Berson, M. J. (2021). Machine Learning and the Social Studies. *Social Education*, **85**(2). pp. 88-92. DOI: https://www.socialstudies.org/social-education/85/2/machine-learning-and-social-studies.

Lee, B. C. G., Mears, J., Jakeway, E., Ferriter, M., Adams, C., Yarasavage, N., Thomas, D., Zwaard, K. and Weld, D. S. (2020). The Newspaper Navigator Dataset: Extracting Headlines and Visual Content from 16 Million Historic Newspaper Pages in Chronicling America. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. (CIKM '20). New York, NY, USA: Association for Computing Machinery, pp. 3055–62. DOI: 10.1145/3340531.3412767.

Lee, B. C. G. and Weld, D. S. (2020). Newspaper Navigator: Open Faceted Search for 1.5 Million Images. *Adjunct Publication of the 33rd Annual ACM Symposium on User Interface Software and Technology*. (UIST '20 Adjunct). New York, NY, USA: Association for Computing Machinery, pp. 120–22 DOI: 10.1145/3379350.3416143.

Piper, A., Wellmon, C. and Cheriet, M. (2020). The Page Image: Towards a Visual History of Digital Documents. *Book History*, **23**(1). Johns Hopkins University Press: 365–97. DOI: 10.1353/bh.2020.0010.

Wevers, M. and Smits, T. (2020). The Visual Digital Turn: Using Neural Networks to Study Historical Images. *Digital Scholarship in the Humanities*, **35**(1): 194–207 DOI: 10.1093/llc/fqy085.

Between Interactive Fictions and Visual Novels: Diversity of Agency in Videoludic Novels

Lescouet, Emmanuelle

emmanuelle.lescouet@umontreal.ca
Université de Montréal, Canada

Dumoulin, Pierre Gabriel

dumoulin.pierre_gabriel@courrier.uqam.ca
Université du Québec à Montréal

Interactive fictions are digital narrative artworks that require little to no dexterity. These multimedia fictions do not require a particular proficiency in their manipulation to access and explore the story. However, they demand distinct mechanics and reading moves, specific to narrative models (Juul, 2007). These items, between literature and videogame studies, call for the analyses of their own modes of interaction and consumption (Aarseth, 1997; Hayles, 2007). Reading and book theories will be of use in this presentation, as we consider the latter as a form of the ludic, notably through the establishment and transgression of "magic circles" (Huizinga, 1938; Picard, 1986; Macé, 2011) and of game studies (Triclot, 2017).

In this specific environment, the reading is one of entertainment, voluntary and engaged. Since there is a tacit contract that readers sign with the work, establishing a horizon of expectation and conveying reading habits (Jauss, 1978; Iser, 1976), our study establishes the purpose of this reading experience as immersion, plunging into a second world, and exploring the possibilities offered by it. Following Ryan (2001), we define the immersion as an entertainment experience where the human subject gets lost inside the text. Xe is transported into the fictional universe; when xe is incorporated (Calleja, 2013), reproducing reading moves and interacting with the narrative mechanics. Not that fiction and reality are confused, but rather that the reader finds themself in the narrative and their potential affects (Massumi, 2015).

We must therefore consider a continuum between artworks relying on a few interactions and others that require more complex sequences, in constrained times. This axis allows us, on the one hand, to recognize the complexity of narrative practices and, on the other hand, to establish anchor points throughout the continuum without basing our research strictly on opposite relationships.

Since the physical inscription in the cultural work and the sensations which result from it strongly influence the reception - the readings - of these works (Citton, 2012; Bigé, 2018; Garmon, 2020), the body's implication in these gestures poses the question of the place of the work, an implication located between the body of the player, the game and the space between them. This varied involvement, created by the different reading gestures, modifies both the narrative mechanics and the immersion possibilities (Ryan, 2004; Murray, 1997). Understanding the implications of these mechanics is necessary to approach a theory of

novel mechanics in game novels. Furthermore, the objective of this study has been to move towards an analysis of the geographical and historical shifts of these forms.
If visual novels are a digital form strongly anchored in Asian digital culture, Western creations and adaptations take up and adapt narrative mechanics including them in more typically Western videogame structures, such as the Quick Time Events. The importance of Asian cultures in the corpus and communities of Western reception leads us to this comparative study in order to understand the implications of these formal intercultural transfers.
These reflections have focused both on a corpus of Japanese and Western interactive fictions, but also on a segmentation of their characteristics: works employing Quick Time Events, according to the expression of the developer Yu Suzuki, such as Detroit: Become Human, Robotics;Notes; games that require minimal interaction : Synergia and Doki Doki Literature Club!; and games that alternate dialogue phases with certain mini-games, which influence the development of narrative arcs: Coffee Talk and Root Letter.
This presentation will begin by describing the reading gestures and the narrative mechanics, before studying the continuum mentioned above and thus approaching the transfers that have taken place between the cultural areas. Through the study of the three categories of interactive fictions, we situate the reading gestures and the narrative mechanics, following the studios' practices. From the establishment of this comparative approach based on Western and Eastern practices, we then analyse a small selection of the corpus as to determine how narrative traditions evolve in the video game universe, but more specifically to identify the main components of both Western's and Japanese's influence. These components then allow us to draw conclusions on how videogame practices shape the development of interactive fictions and the reader/player's experience. The comparative corpus is based on the ontologies of digital works from the Canada Research Chair in Digital Textualities Repository, which serves as a tool for the corpus construction and statistical analyses necessary for this presentation.

Bibliography

Aarseth, E. (1997) *Cybertext: Perspectives on Ergodic Literature*. Baltimore: Johns Hopkins University Press.

Bigé, R. (2018) 'Note sur le concept de geste', *Pour un atlas des figues*. Available at: https://hal.archives-ouvertes.fr/hal-02440249/document (Accessed: 27 July 2021).

Calleja, G. (2011) *In-Game: From immersion to Incorporation*. Londres: MIT Press.

Citton, Y. (2012) *Gestes d'humanités: anthropologie sauvage de nos expériences esthétiques*. Paris: Armand Colin (Collection: Le temps des idees).

Garmon, I. (2020) 'Le corps à l'épreuve des applications : des « petits gestes » éprouvants ?', *Les Chantiers de la Création*, La mise à l'épreuve du corps(12). doi:https://doi.org/10.4000/lcc.3102.

Hayles, K.N. (2007) *Electronic literature: what is it?* Available at: https://eliterature.org/pad/elp.html (Accessed: 2 April 2019).

Huizinga, J. (1938) *Homo ludens. Essai sur la fonction sociale du jeu*. Translated by C. Sérésia. Paris: Gallimard (Les Essais).

Iser, W. (1976) *L'acte de lecture: Théorie de l'effet esthétique*. Translated by E. Sznycer. Pierre Mardaga.

Jauss, H.R. (1978) *Pour une esthétique de la réception*. Translated by C. Maillard. Paris: Gallimard (Bibliothèque des Idées).

Juul, J. (2005) 'Fiction', in *Half-Real: Video Games Between Real Rules and Fictional Worlds*. Cambridge: MIT Press, pp. 121–162.

Macé, M. (2011) *Façons de lire, manières d'être*. Paris: Gallimard (NRF essais).

Massumi, B. (2015) *Politics of Affect*. Wiley.

Murray, J.H. (1997) *Hamlet on the Holodeck: The Future of Narrative in Cyberspace*. Cambridge, MA: MIT Press.

Picard, M. (1986) *La lecture comme jeu : essai sur la littérature*. Paris: Les Éditions de Minuit (Critique).

Ryan, M.-L. (2001) 'the Text as World', in *Narrative as Virtual Reality. Immersion and Interactivity in Literature and Electronic Media*. Baltimore: John Hopkins University Press, pp. 89–114.

Ryan, M.-L. (ed.) (2004) *Narrative across media: the languages of storytelling*. Lincoln: University of Nebraska Press.

Triclot, M. (2017) 'Play studies', in *Philosophie des jeux vidéo*. Paris: La Découverte, pp. 15–42.

Gender and Cultural Diversity in Chinese Children's Picture Books: A Data-led Analysis of Bestselling Modern Titles

Li, Yi
s2127837@ed.ac.uk
School of Literatures, Languages & Cultures, University of Edinburgh, United Kingdom

Terras, Melissa

m.terras@ed.ac.uk
College of Arts, Humanities and Social Sciences, University of Edinburgh, United Kingdom

Li, Yongning
lee9512@163.com
School of Systems Science, Beijing Normal University

Introduction

Picture books are a main source for pre-schoolers to learn about the wider world; it is important for children to see themselves in books and be aware of differences (Johnston and Bainbridge, 2017; Latima, 2020). Male dominance in children's books, such as the dominance of male characters has long been problematic (Gooden and Gooden, 2001; Kim, 2016; Terras, 2018), it is important therefore to consider gender (as a protected characteristic) when considering diversity in the Chinese children's book market.

Picture books in different countries reflect diverse cultural preferences (Saxby & Winch, 1987; Wee et al., 2015). With a large, growing children's picture book market (Johnson, 2018), China has translated children's titles from countries including US, UK, Japan, etc (Li et al., 2020). This study examines the diversity in gender and popular themes in Chinese children's picture books, by analysing the book authorship, titles, and blurbs of 2,000 bestselling children's picture books from Dangdang, the major Chinese online bookseller. It provides a general reflection of topics in children's books from different countries, including Asian, Western countries and other regions.

Method

Data Corpus

We conducted an experimental approach from publicly available information online. Metadata from 2,000 best-selling pre-school picture books were scraped from Dangdang.com, using Python. All data, including book title, blurbs, author introductions were collected on 24 th September 2020 and filtered by sales, the earliest title was published in 2003. Data was allocated to four separate corpora regarding original language and region of publication including Chinese local, East Asian (Japan, South Korea, etc), English (US, UK, Canada, etc) European, and multiregional books (French, Germany, South Africa, etc).

Data Analysis

This study firstly matched all authors with their sex and nationality by analysing authors' introductions, as well as searching for official authorial information online, then compiling a statistical breakdown. Secondly, we tokenized book titles and blurbs using the Jieba package in Python, calculated the top 1,000 word list for each corpus. We then identified gendered words, classified them into five groups: pronouns (he/she); gender roles (mum/dad); nouns (princess/witch); animals (cow/cock); name of character (Carmela/Tintin); calculated frequency and compared them.

Thirdly, we adopted topic modelling, a text mining method for understanding contents of a corpus through a group of topics (Heidarysafa et al., 2019), to investigate how book topics differ. We used Bertopic, which supports English and Chinese, and tested different parameters for each corpus. We finally modified the number of topics to 3 and 30 topic words, with a value for each word showing its centrality to the topic.

All data was collected in Chinese with Chinese-style narratives. We analysed data in Chinese then translated the results into English, as presented below. Finally, we tried to match characters' names with their original versions and present English names (if there were any).

Results

Comparison of Authors

China had imported picture books from 34 countries and areas, with a male dominance among authors (see in Fig1&2). There are 519 East Asian titles, 730 English titles and 332 multilingual titles.

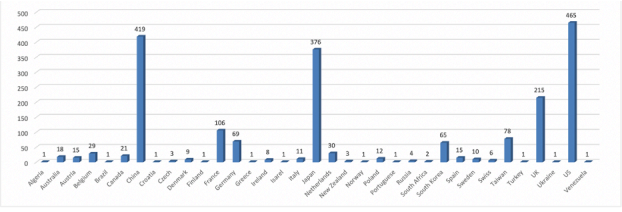


Fig 1
Nationality of Authors

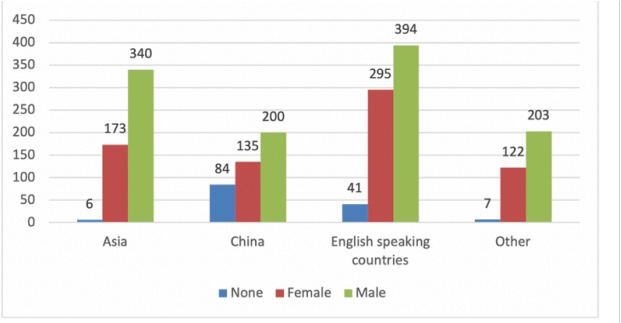


Fig 2
Gender of Authors 1

Gender representation in books

A gendered keyword list compared the gender representation in books. There are more male characters in all corpora, however, mum/mom weighed more than dad in all books. Chinese and Asian titles have fewer gendered words, while more characters were portrayed in English books, correspondingly more pronouns are used.

Table1 Gendered Keywords Frequency

Gendered words (number of gendered word and word frequency)	China		East Asia		English Speaking Countries		Other	
	Male	Female	Male	Female	Male	Female	Male	Female
Pronoun (He/she etc.)	1 (244)	1 (130)	1 (407)	1 (174)	1 (1434)	2 (770)	2 (786)	1 (313)
Gender roles (Dad/mom etc.)	3 (106)	2 (154)	9 (542)	7 (509)	6 (499)	6 (753)	4 (242)	6 (295)
Noun (King/princess etc.)	7 (207)	3 (67)	8 (303)	4 (99)	3 (345)	6 (281)	8 (139)	5 (73)
Animals (Cock/cow/hen etc.)	1 (10)	/	1 (32)	/	/	1 (18)	/	3 (79)
Character (Tintin/Nicky etc.)	11 (134)	4 (67)	4 (104)	2 (25)	15 (788)	12 (418)	29 (523)	14 (267)
Total	701	418	1388	807	2955	2260	1706	1112

* Numbers in every blank respectively represent the number of different words in this type and the total frequency of all words in this type.

Topic clustering between corpora of different regions

Topics from each dataset were presented in word clouds (Fig 3-14). Chinese culture, children's education and books resulting from corporate franchises are three main topics in Chinese local titles. Similar topics can be found in East Asian titles, but there are more local topics. Big brands such as Disney, family education and animal stories are important aspects in English titles.



Fig 3-5 Topics in Chinese Titles

Fig 6-8 Topics in East Asian Titles

Fig 9-11 Topics in English Language Titles

Fig 12-14 Topics in Multilingual Titles

Discussion

The data-driven analysis of book authors and blurbs indicates publishing, purchasing, and reading preferences and showed an overall male dominance the Chinese children's picture book market. Chinese and East Asian titles emphasise cultural contents and are more education-oriented, while books from Western countries portray more characters and focus on storytelling and children's emotions. However, this study is a partial reflection of the Chinese children's book market due to limited data collected for popular titles. All translated titles were chosen under the publishing censorship in China, and they might not be proper representatives of diverse picture books. Further

studies can expand the dataset, include more languages and book genres, as well as adopting other methods such as network analysis.

Bibliography

Gooden, A. M. and Gooden, M. A. (2001). Gender Representation in Notable Children's Picture Books: 1995-1999. *Sex Roles*, **45**(1/2), pp. 89–101. http://dx.doi.org/10.1023/A:1013064418674.

Johnston. and Bainbridge, J. (2017). *Reading Diversity through Canadian Picture Books: Preservice Teachers Explore Issues of Identity, Ideology, and Pedagogy*. University of Toronto Press. 10.3138/9781442666412.

Johnson. (2018a). *China's Children's Book Market: Big Numbers and Local Talent*. *Publishing Perspectives*. https://publishingperspectives.com/2018/11/china-childrens-book-market-big-numbers-local-talent/ (accessed 25 October 2021).

Johnson. (2018b). *Why Children's Book Publishing in China Is Growing So Fast*. *Publishing Perspectives*. https://publishingperspectives.com/2018/03/why-chinas-childrens-book-industry-is-growing-so-fast/ (accessed 25 October 2021).

Kim, S. J. (2016). Expanding the Horizons for Critical Literacy in a Bilingual Preschool Classroom: Children's Responses in Discussions with Gender-Themed Picture Books. *International Journal of Early Childhood*, **48**(3), pp. 311–27. http://dx.doi.org/10.1007/s13158-016-0171-3.

Li, Yi., Hangxizi, S. and Li, Yongning. (2020). Is the Chinese Children's Mainstream Book Market Inclusive Enough? A Data Analysis of Children's Bestsellers on Dangdang.Com. *Publishing Research Quarterly*, **36**(1), pp. 129–44. http://dx.doi.org/10.1007/s12109-019-09706-z.

Saxby, H. M. and Winch, G. (1987). *Give Them Wings: The Experience of Children's Literature*. South Melbourne: Macmillan Co. of Australia.

Wee, S.-J., Park, S. and Choi, J. S. (2015). Korean Culture as Portrayed in Young Children's Picture Books: The Pursuit of Cultural Authenticity. *Children's Literature in Education*, **46**(1), pp. 70–87. 10.1007/s10583-014-9224-0.

Notes

1. There are some titles were written by a group of editors, it is difficult to confirm their sex information, so we classified them into the none group.

DH Pedagogies in the Global Souths

Licastro, Amanda Marie
amanda.licastro@gmail.com
The University of Pennsylvania, United States of America

Roy, Dibyadyuti
dibyadyutir@gmail.com
Indian Institute of Technology (IIT) Jodhpur

Esprit, Schuyler Kirshten
schuyleresprit@gmail.com
Create Caribbean Research Institute, Dominica

This presentation will be given by the editors of a special issue of *Reviews in DH* focused on Digital Humanities Pedagogy in the global context. Inspired by recent efforts to validate DH pedagogy through formal publication, such as the edited collection *Digital Pedagogy in the Humanities*, and journals such as *Hybrid Pedagogy* and *The Journal of Interactive Technology and Pedagogy,* we aim to bring awareness to the rich content created by students in higher education. In this presentation we will highlight the projects reviewed in the special issue, and we intend to include the voices of both the creators and the reviewers in order to make transparent the labor involved in building this publication. We hope to acknowledge the shape and infrastructure of pedagogy in the context of Global South(s) that often fall outside the normative forms reflected in these US-centric journals.

In this issue we focus on projects that demonstrate the power of community-engaged, research-based digital humanities across contexts. The intention is to showcase a range, both in terms of tools and scale, as well as a diversity in the production process from a variety of institutions across the globe. Some of the projects in this issue deploy open source and open access digital tools for project building such as Omeka and WordPress. These are accessible for most undergraduate students entering the field while also being accessible in the Global South where proprietary tools are cost prohibitive even when they are available within regions outside of the developed world. Participants adding content and updating infrastructure to sustain these digital spaces carefully over time raises questions about project sustainability. Given the realities of the Covid 19 pandemic, climate vulnerability, and the onslaught of disasters that threaten the Global South from the Caribbean to South Asia, we hope that these projects

inspire creators and future DH teachers and practitioners to think more deliberately about preserving the digital record.

Being acutely aware that DH and its pedagogy(ies) are shaped by the infrastructures, limitations, and affordances of our local contexts, we highlight projects that exemplify the chaotic yet productive potential of digitality, while also being alive to its various discontents. We believe that this special issue of *Reviews in DH* exemplifies how DH pedagogy must at its core focus on two key facets: empathy and engagement. With a keen eye toward the relational and not only geographical definitions of Global DH, this issue eschews authoritatively defining "DH Pedagogy." Instead, the projects are illustrative of the complex genealogies and overlaps between cultural connotations of digital pedagogy, often made invisible in normative DH conversations.

The richness of the student projects in this issue has pushed the presenters to call for illuminating the often-invisible labor of digital humanities project development, even within resource-rich institutions. Many of the critiques raised in the reviews found in this journal issue, could be potentially addressed through a page on process in these projects, or a blog of reflection where project creators can lay bare the experimental and transformative nature of DH work in these specific circumstances. Perhaps this self-reflexive model of project development may become best practice through course sites where both instructors and students think through work together, or even with the inclusion of course syllabi as a menu option, elucidating the intellectual framework within which iterative student learnings emerge.

In this presentation we will walk the audience through the process of working on this special issue across international time zones. We will discuss the main points of synchronicity and tension that arose, as well as the challenges and rewards of identifying projects and reviewers that would accurately represent the wide range of possibilities to inspire pedagogues working around the globe.

Rethinking the Advanced Research Consortium: Disciplinary Restructuring and Linked Open Data

Liebe, Lauren
leliebe@tamu.edu
Texas A&M University, United States of America

Mandell, Laura
mandell@tamu.edu
Texas A&M University, United States of America

The Advanced Research Consortium (ar-c.org) is a hub of humanities research nodes focused on aggregating and peer reviewing digital archives, collections, and research resources. Each node hosts an online "finding aid" which serves as a centralized research space for exploring traditional scholarship such as journal articles, digital collections (Early English Books Online and Eighteenth-Century Collections Online, for example), and scholarly digital resources that ARC peer-reviews, using the same process as journals and university publishers in the humanities. These nodes aggregate millions of digital artifacts from across the digital humanities spectrum, from proprietary projects such as JSTOR and Adam Matthew Digital to independent digital humanities projects like the London Stage Database and the Lili Elbe Digital Archive, as well as collections from major libraries throughout North America and Europe.

ARC has a long history. Its roots lie in Jerome McGann's founding of NINES (Networked Infrastructure for Nineteenth-Century Electronic Scholarship) in 2003. The initial Steering Committee was led by Jerome McGann and Bethany Nowviskie and included Morris Eaves, Neil Fraistat, Steven Jones, Laura Mandell, Kenneth Price, and Martha Nell Smith, all of whom had created digital archives for nineteenth-century studies. Their resources would be peer-reviewed and made findable, at the level of objects in the archive, through NINES.org. The goal was to de-silo these projects, to render them all searchable alongside proprietary resources. NINES fit its nodes and peer review process into traditional humanities field structures to render the peer-review process legible to traditional colleagues: renowned experts served on the editorial boards, rendering them as illustrious as any board for a university press. Through this merging of traditional scholarly apparatus with ever-evolving digital work, NINES, and later ARC, translated digital humanities work into familiar structures for tenure and promotion, and made them visible for teaching and research.

ARC is a response to the need for aggregating and providing peer review to communities of scholars beyond those specializing in nineteenth-century studies; it supported the development of medieval, eighteenth-century, and modernist nodes. Later in ARC's history, thanks to the efforts of Michigan State University, ARC began creating nodes for libraries with special collections that wanted to make their holdings searchable alongside relevant archives and journal publications. MSU's Studies in Radicalism (SiRO) led the way in rethinking ARC's traditional, period-centric structure.

While conservative on the front end, ARC's original backend was technologically advanced, thanks to the prescience of (now) Dean Nowviskie: [1] the metadata was rdf xml in format which was loaded into SoLR with a Lucene search engine (Nowviskie, 2007). This metadata format (Submitting RDF) made it possible to add URIs to create Linked Open Data and for ARC to organize search returns beyond traditional disciplines.

ARC was launched when content management systems and cultural heritage platforms were in their infancy. To maintain robust and flexible metadata and search capabilities, we are sunsetting our legacy COLLEX software and moving to WordPress to allow nodes more control over their web presence. These WordPress instances will interface with an updated backend: the Corpora Dataset Studio, developed by ARC technology director Bryan Tarpley. Corpora uses mongoDB as well as elasticsearch and neo4j databases and has been designed with data visualization and Linked Open Data in mind. In collaboration with Linked Infrastructure for Networked Cultural Scholarship (LINCS, Susan Brown, PI), Tarpley is recreating ARC's legacy visual search interface, BigDIVA, in the form of a Rich Prospect Linked Open Data Viewer. Susan Brown has argued that Linked Open Data can demonstrate that categories such as disciplines, gender, and nationality can be revealed as "categorically provisional" through forging crosswalks between traditional categories and new, anti-disciplinary, intersectional, decolonizing, and queer modes of categorization (Brown, 2020). The new LOD viewer will allow users to dynamically organize the ARC catalog in their own ways, using categories and queries that make sense to them.

ARC is now working with special interest groups – from the American Antiquarian Society to the Canadian Writing Research Collaboratory and Escalator in South Africa – to create new nodes, and we are actively recruiting nodes and peer reviewing projects that reconsider the boundaries of traditional humanities subjects. ARC seeks partnerships with other groups who are in the same situation as the original NINES steering committee: they would like their projects to be searchable along with other digital projects and resources. To this end, we are developing nodes focused on disability studies, early modern drama, music studies, and more.

If ARC can support such groups, it will live up to that very tendentious but invigorating hope expressed in the original *Digital Humanities Manifesto* composed by the DH group at UCLA: " Traditional Humanities is balkanized by nation, language, method, and media…. [W]e imagine different constellations (not just disciplinary constellations, but also other configurations of producing knowledge that can be team- and project-based, collaborative, open-ended, globally-oriented, engaging for new audiences and institutions)"(Digital Humanities Manifesto, 2008). [2] ARC and Corpora – to be released open source next year, and usable on a laptop – aspire to constitute a "diversity stack" (Liu, 2020)

Establishing nodes related to special communities and interests can constellate the scholarly research universe in un- or even anti-disciplinary ways. Scholars everywhere will be able to get involved in these communities, to join the editorial boards, contribute projects for peer review, attend ARC workshops, etc. Insofar as "de-disciplining" requires de-colonizing as well, ARC is expanding Corpora's capacities to read non-Latin script and make it possible to better accommodate "expanding the notion of evidence" to include "fragments of events, depicted in periodicals, testimony, pamphlets and even poetry, in the archive" (Risam, 2015).

The authors present these two modes of de-disciplining ARC, viz., 1) supporting emergent (anti-)disciplines in co-developing an infrastructure free of colonial archiving constraints, from the ground up; and 2) transforming the ARC catalog from Enlightenment-style index (Pasanek and Wellmon, 2015) into a Linked Open Data viewer organized by SPARQ-L queries, in order to elicit ideas from the audience about how ARC can further support spontaneous constellations of intellectual community that may contribute to the evolution of university disciplines in the Humanities.

Bibliography

Brown, S. (2020). Categorically provisional. *PMLA*, 135(1): 165-174.

Liu, A. (2020). Toward a diversity stack: Digital humanities and diversity as technical problem. *PMLA*, 135(1): 130-151.

Nowviskie, B. (2007). A scholar's guide to research, collaboration, and publication in NINES. *Romanticism and Victorianism on the Net*, https://doi.org/10.7202/016707ar.

Pasanek, B. and Wellmon, C. (2015). The Enlightenment index. *The Eighteenth Century*, 56(3): 359-382.

Risam, R. (2015). Revising history and re-authouring the left in the postcolonial digital archive. *Left History* 18(2): 35-46.

Notes

1. See also Jerome McGann, Bethany Nowviskie, "NINES: a federated model for integrating digital scholarship", *Electronic Book Review*, 31 January 2012, an earlier form of which is available here: https://nines.org/about/wp-content/uploads/2011/12/9swhitepaper.pdf.

2. The Digital Humanities Manifesto 2.0, dated 5/29/2009, is described and available for download: http://www.toddpresner.com/?p=7. The original manifesto, dated 12/15/2008, is no longer available: email mandell at tamu dot edu for a saved copy.

Discovering Civil Disputes Hidden in the Court Judgment Documents for Applications in Social Studies and Legal Informatics

Liu, Chao-Lin
chaolin@g.nccu.edu.tw
National Chengchi University, Taiwan

Liu, Yi-Fan
108753213@nccu.edu.tw
National Chengchi University, Taiwan

Liu, Wei-Zhi
109753157@nccu.edu.tw
National Chengchi University, Taiwan

Lin, Hong-Ren
109753156@nccu.edu.tw
National Chengchi University, Taiwan

Background

Civil and criminal cases provide direct clues about the relatively serious conflicts in a human society. These conflicts were not resolvable privately or easily via mediation, so were resorted to the legal system. The judgment documents publicized by the courts offer opportunities for us to analyze the main causes of the conflicts, and we focus on the civil disputes in this study.

The analysis of civil cases and the resulting observations are relevant to multiple social issues. Researchers studied the lawsuits for litigation strategies (Huang et al., 2010). One may study the previous cases to understand whether mediation will be a better choice than litigation (Anderson et al., 2018). Analogous research procedure that relied on the Chinese local gazetteers may help us investigate the social conflicts in the history (Li, 2000).

We aim at understanding the main causes that led to the civil disputes, and report preliminary results of analyzing the cases of labor and employment (**L&E**, henceforth) litigations and of family support (**FS**, henceforth) litigations. This line of work is possible because the judicial law in Taiwan requires the courts to publicize their judgment documents except for special cases with specific concerns and because we can obtain and analyze these open data with computing techniques.

We show the main steps of our work in Figure 1.

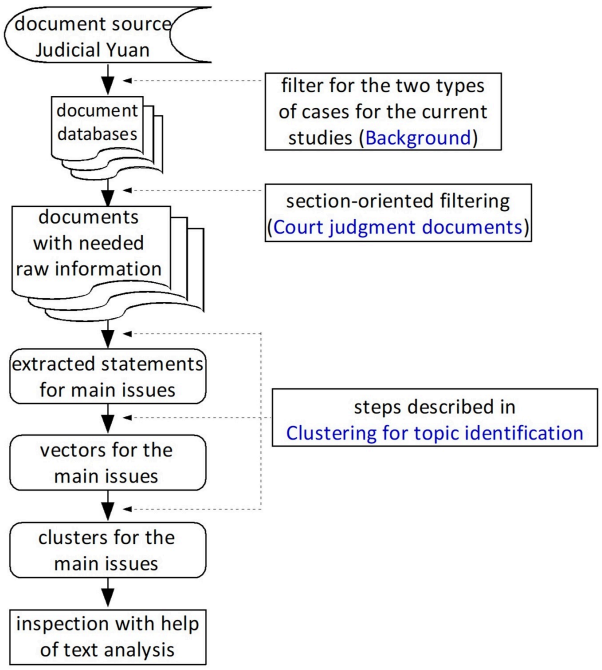


Figure 1.
Main steps of our work

Court judgment documents

We may download the court judgment documents from the official website that is maintained by the Judicial Yuan which governs the courts in Taiwan. [1] For the L&E litigations, we use 3524 cases of between 2007 and 2020. There are three tiers of courts, and we use only the documents of the lowest tier, i.e., the local courts. For the FS litigations, we use 1223 cases of between 2014 and 2020 and also of the local courts.

There were a lot more relevant litigations than the number of documents that we use in the current studies. We chose documents of the local courts because the documents usually provide more preliminary and direct information about the disputes. In addition, only a portion of the documents for the L&E litigations contain standard segments that should record a summary about the specific

issues that the plaintiffs and the defendants were contesting. Without such specific summaries, it is not easy to identify the issues being contested, which were recorded in typically long judgment documents. Hence, we chose only 3524 instances that provide such summaries. Notice that, even if we have the summaries, it is not easy for computer programs to "understand" the statements. The statements were written for individual cases in the form of natural language (not keywords, for instance), and the wordings often differ for conflicts of the same type.

The documents for the FS litigations may have paragraphs that provide information about the reasons that the plaintiffs were seeking for financial assistance from the defendants and about the reasons that the defendants did not agree with the requests. We employed a keywords-based approach to identify 1223 documents that have such paragraphs for the current studies. Similar to what we discussed above, reasons of the same type might be written in very different styles in the documents.

Clustering for dispute identification

While we selected the documents with specific sections, we also extracted the statements within those sections. As we just explained above, we believe that these statements described the disputes in question. We tried some different ways to preprocess the extracted statements, which consisted of typical steps for natural language processing, but we could not explain them clearly in this abstract. We obtained more than 17000 statements and 1223 paragraphs for the L&E and FS litigations, respectively.

We then converted the statements and paragraphs to vectors using both the typical TF-IDF vector-space-model approach (Croft et al., 2010) and the SBERT pretrained model (Reimers and Gurevych, 2019). In computer science, we hope that the vectors somehow represent the semantics of the original statements.

We apply and hope that clustering (Alpaydin, 2020) the vectors of statements would lead us to identify different types of issues. For this step, we explored the well-known *k*-means clustering [2] and the FAISS method of Facebook (Johnson et al., 2017). In essence, we are applying the concept of topic modeling with tools of our choice (Blei, 2010a, 2010b).

Observations and discussion

The results of topic modeling can be useful if we inspect and interpret the results with some appropriate principles (Ramage et al., 2009; Sievert and Shirley, 2014). For the L&E litigations, we could identify interesting topics that were indicated by the statements about issues in individual clusters. Inspection by human experts certainly takes time, but that is still more efficient than reading the complete judgment documents over the years directly. The disputes may be caused by different types of compensation/payment problems. Some examples follow.

1. for retirement benefits
2. for unlawful or debatable layoff
3. for body injuries or fatality during worktime
4. for the salaries and the late-night meals as a result of overtime work

Some clusters contain statements about non-monetary issues or for special types of labor force.

1. the interpretation of the non-compete clause
2. the time for paid leave
3. the disputes for sailors

Once we identify these topics, we can find the lawsuits that share the same or similar combinations of disputes. Since we know when and which local courts handled these litigations, we can analyze and visualize such similar cases spatiotemporally, offering a foundation for social studies. We can also build focused databases by collecting information about similar litigations, thus providing references for future judgments.

In contrast, our current achievements for the FS litigations were less impressive. Our clustering approach could find that the involvement of the parents of the couple that was fighting for divorce. We could also algorithmically identify cases in which adult children battled for fair shares for the support of their retired parents. The issues behind FS litigations often consist of complex and mixed daily problems, and our current algorithms could not differentiate the core disputes precisely yet. We report the relatively poor results for the FS litigations to contrast our promising accomplishments for the L&E litigations in this proposal, and we certainly will continue to refine our approach for the FS litigations.

We are thankful to the reviewers, and shall provide more technical details about our work in the oral presentation as requested. This work was supported in part by the Ministry of Science and Technology of Taiwan.

Notes

1. The website of the Judicial Yuan is located at <https://opendata.judicial.gov.tw/>

2. The software tools were implemented in scikit-learn and accessible at <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>

Bibliography

Alpaydin, E. (2020). *Introduction to Machine Learning*, fourth edition, chapter seven, The MIT Press. (Clustering, a type of machine learning technique, aims to put similar data into a cluster. Given a collection of data, clustering may assign individual datum into one of a selected number of clusters.)

Anderson, D. Q., Chua, E. and My, N. T. (2018). How Should the Courts Know Whether Dispute Is Ready and Suitable for Mediation: An Empirical Analysis of The Singapore Courts' Referral of Civil Disputes to Mediation, *Harvard Negotiation Law Review*, 23(2):265–318.

Blei, D. M. (2012a). Probabilistic Topic Models, *Communications of the ACM*, 55(4):77–84.

Blei, D. M. (2012b). Topic Modeling and Digital Humanities, *Journal of Digital Humanities*, 2(1). <http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/>

Croft, W. B., Metzler, D. and Strohman, T. (2010). *Search Engines: Information Retrieval in Practice*, 241–247, Pearson.

Huang, K.-C., Chen, K.-P. and Lin, C.-C. (2010). An Empirical Investigation of Settlement and Litigation—The Case of Taiwanese Labor Disputes, *Journal of Empirical Legal Studies*, 7(4):786–810.

Johnson, J., Douze, M. and Jégou, H. (2017). Billion-Scale Similarity Search with GPUs, arXiv preprint, arXiv:1702.08734. <https://github.com/facebookresearch/faiss>

Li, B. (2000). Civil Disputes and Law Suits in Huizhou of The Ming Dynasty, *Historical Research*, Chinese Academy of History, I:94–105.

Ramage, D., Rosen, E., Chuang, J., Manning, C. D. and McFarland, D. A. (2009). Topic Modeling for the Social Sciences, *Proceedings of the Workshop on Applications for Topic Models: Text and Beyond*, Twenty-Third Conference on Neural Information Processing Systems.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the Ninth International Joint Conference on Natural Language Processing*, 3982–3992. <https://www.sbert.net/>

Sievert, C. and Shirley, K. E. (2014). LDAvis: A Method for Visualizing and Interpreting Topics, *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63–70.

Digital Resource Aggregation: Giving New Life to Multi-source Cultural Data

Liu, Rui
ruiliu2@student.unimelb.edu.au
The University of Melbourne, Australia

McKay, Dana
dana.mckay@rmit.edu.au
Royal Melbourne Institute of Technology, Australia

Buchanan, George
george.buchanan@unimelb.edu.au
The University of Melbourne, Australia

Digital humanities have traditionally been concerned with utilizing digital technologies in processing cultural resources such as text, images, and specialist data (e.g. geographic information) (Tiepmar, 2018). In practice, many digital humanities projects aggregate cultural data from multiple sources. There are two types of aggregation, where the first one collects data from original (unpublished) datasets to form a new digital humanities collection, while the second integrates parts from existing collections by sharing data or forming a new system from them (Siqueira and Martins, 2021, Freire et al., 2018). This research explores both forms of digital resource aggregation in digital humanities and illustrates how aggregation breathes new life into multisource cultural data. Taking the perspective of the construction of digital humanities collections, our case studies discuss different stakeholders' views of digital resource aggregation and their work practices. The following three research questions will be addressed in this paper:

1. What are the digital resource aggregation approaches of digital humanities collections?
2. What are the problems encountered when aggregating digital resources?
3. What are the key lessons learned when attempting digital resource aggregation?

Our research applies semi-structured individual interviews to understand the practical experience of digital humanities projects that brought content together from multiple existing collections. The data were collected from

September to November 2021. We recruited two pilot study and nineteen main-study participants. All participants are digital humanities scholars active in major research centres and conferences with digital humanities projects. During interviews, we asked participants about their background, the details about their data aggregation project, their role in the project, the aggregation method they used, aggregation problems, their advice and perspective of multisource cultural data aggregation through digital humanities collections. Each interview took approximately 60 minutes and all interviews were recorded and transcribed. We coded our data based on the three research questions (shown above) using NVivo. Eleven main participants have direct experience of data aggregation. Followed by that, we formed these experiences into case studies of individual projects.

Our paper focuses on the eleven cases of integrating multisource cultural data. Each case study considers a single project's aims, team, construction steps, digital resource aggregation approaches, and any suggestions from interviewees for how to overcome challenges. The aggregation datasets in these eleven case studies include medieval manuscripts and illuminated manuscripts, religious and political documents, digital archival resources for medieval to modern history, biography of historical people, places information, archaeological data sets, music data sets and annotation works. Our participants included two developers, one Ph.D. candidate with good programming skills, one librarian who acquires programming by self-learning, one student research assistant, and six project managers with socio-technical background.

These eleven cases represent a wide variety of different approaches. Some use more than one: five cases use Linked Data to integrate multisource data, three of them use the International Image Interoperability Framework (IIIF) to integrate images, two use a CMS system to manage data, one uses a SQL database to aggregate different data, one uses an API to link different digital humanities collections, one uses XML to do digital resource aggregation and one makes metadata protocols to do aggregation.

The results of our case-studies reveal different digital resource aggregation challenges and solutions within digital humanities, and suggest there may be untapped possibilities for aggregation in digital humanities research. Key problems in digital resource aggregation can be categorized into four aspects, namely data problems, system problems, institution problems and skillset problems.

The data problems are related in (1) some unstructured data, such as ambiguous metadata schemata and uncertain standards, lack of digital readiness, and issues of different data formats; (2) integrating data with the same type but that have different data levels of detail, e.g.time, and (3) original data sets that cannot be shown in the aggregator website would make aggregation of digital humanities collections harder. For system problems,the aggregation of different data sets would be influenced by system updates, software changes, systems interaction issues and visualization problems. For institution problems, funding and licensing are important. For skillset problems, the distribution of skills among digital humanities researchers and technical threshold of aggregation technology raises the required time and labor for higher aggregation performance.

The digital humanists we interviewed provided us valuable advice when attempting to achieve digital resource aggregation, such as unifying metadata standards, improving data sharing policy and data explanation, encouraging collaboration from community and administrators, and enhancing digital humanities research infrastructure and project sustainability.

Bibliography

FREIRE, N., MEIJERS, E., VOORBURG, R. & ISAAC, A. 2018. Aggregation of cultural heritage datasets through the Web of Data. *Procedia Computer Science,* 137 **,** 120-126.

SIQUEIRA, J. & MARTINS, D. L. 2021. Workflow models for aggregating cultural heritage data on the web: A systematic literature review. *Journal of the Association for Information Science and Technology*.

TIEPMAR, J. 2018. Big Data and Digital Humanities. *Archives of Data Science, Series A,* 5.

Mining the Native American Authored Works in HathiTrust for Insights

Lu, Kun
kunlu@ou.edu
University of Oklahoma, United States of America

Heaton, Raina
rainaheaton@ou.edu
University of Oklahoma, United States of America

Orr, Raymond
raymond_orr@ou.edu
University of Oklahoma, United States of America

Vetter, Alyssa
alyssavetter@ou.edu
University of Oklahoma, United States of America

Dubnicek, Ryan
rdubnic2@illinois.edu
University of Illinois, United States of America

Magni, Isabella
isamagni@indiana.edu
Indiana University, United States of America

Native Americans represent a historically under-resourced textual community. While there has been an ever-increasing number of Native authors creating works since the 1960s, no corpus of Native-authored works exists from which to draw insights about this particular community, and give them the recognition equal to other similar communities of practice (e.g. History of Black Writing1). In collaboration with the HathiTrust Research Center (HTRC) and with the support of a Scholar-Curated Worksets for Analysis, Re-use and Dissemination grant (SCWAReD2), we have created a preliminary database of Native-authored works, which allows us to use text mining techniques to reveal novel characteristics of this community, such as their identity, worldview, representation, and modes of expression. Text mining also offers a new approach to looking at the ways in which Native authors express themselves and how they may differ from other authors. For example, there is a common assertion that Native peoples forefront the natural world or communitarian relationships more so than non-Native authors (e.g. Schweninger, 1993; Weaver, 1997). Additionally, we are interested in the rhetorical characteristics of the Native American Studies academic genre as illustrated by corpora of Native Studies journals, and how these trends define the evolution of the discipline over time.

Dataset	# of titles searched	# of titles found in HTDL	Coverage
IPL	2013	900	44.71%
NAL	122	37	30.33%
Linguistics dissertations	34	0	0%

Bibliography

Schweninger, L. (1993). Writing Nature: Silko and Native Americans as Nature Writers. *Melus*, **18**(2), 47-60.

Applying LERA for collating witnesses of The Tale of Kiều, a Vietnamese poem written in Nôm script

Luu, Thi Kim Hanh
thi.luu@student.uni-halle.de
Martin Luther University Halle-Wittenberg, Germany

Pöckelmann, Marcus
marcus.poeckelmann@informatik.uni-halle.de
Martin Luther University Halle-Wittenberg, Germany

Ritter, Jörg
joerg.ritter@informatik.uni-halle.de
Martin Luther University Halle-Wittenberg, Germany

Molitor, Paul
paul.molitor@informatik.uni-halle.de
Martin Luther University Halle-Wittenberg, Germany

Introduction

The collation tool LERA [1] is a working environment that combines the entire process of document management, tokenization/segmentation, normalization, alignment and visualization with interactive control options and exploratory tools. It is being successfully applied in several Digital Humanities projects of different languages, e.g., Arabic (Gründler and Pöckelmann 2018, Gründler 2019, Gründler et. al 2020), French (Bremer 2013), Hebrew (Necker et. al 2019, Molitor et. al 2020), Portuguese/ Spanish/Latin (Bragagnolo 2017), English (Alder et. al 2020) and German (Hahn et. al 2020). A comprehensive description of LERA including a comparison with other state-of-the-art approaches for collation can be found in Pöckelmann et. al (2022).

Here we describe a recent experiment in which we applied LERA to the poem Truyện Kiều by Nguyễn Du (1765-1820) to reveal necessary adaptation needs of the software to compare Vietnamese text witnesses written in Nôm script, the historic writing system of Vietnam. Truyện Kiều – or the Tale of Kiều – is one of the most famous texts of Vietnamese literature. For our experiment, we used text witnesses from 1866, 1870, 1871, 1872, and 1902 made available by the Vietnamese Nôm Preservation Foundation (Balaban et. al 2001).

Specifics of Truyện Kiều

The text was written in Nôm script. It follows the traditional Vietnamese verse form Lục bát according to which lines of six Nôm characters (respectively syllables) alternate with lines of eight. The text is organized in pages whose amount of lines varies among the witnesses. The writing direction is top-to-bottom and right-to-left.



Fig. 1:
Facsimile of the first page of Truyện Kiều (1866), adopted from Balaban et. al (2001).

The data includes philological commentaries and a translation of the text in Quốc Ngữ, the modern writing system of Vietnam. There is no one-to-one mapping between the original Nôm script and its translation, although the verse structure is the same in both scripts. This is due to synonyms (e.g., 劄 → chép or ghi), the not normalized placement of diacritics (e.g., 鎖 → khoá or khóa), regional preferences of translators in the choice of words (e.g., 改 → gửi or gởi), and occasional spelling errors (e.g., 色 → sắo instead of sắc).

Data processing

For the processing in LERA, we converted the original plain text files into a XML format according to the Text Encoding Initiative [2] guidelines. Always two lines of Nôm script – one of six, the other of eight characters – were combined into one segment (by <ab>-tag). The original division was preserved with line breaks (<lb/>-tags). We have chosen this approach because two lines normally form a sense unit. Each segment is also extended by its translation. To simplify the tokenization of Nôm script, we added zero-width-spaces (Unicode U+200B) between the individual characters so that lines are not considered falsely as single words by the system. The philological commentaries have been added as well to the XML at their respective positions.

```
1.1 img:/data/kieu/1866/page01a.jpg
1.1 hn: 𥙿𪗋𪚥揆𠊛些
1.1 qn: Trăm năm trong cõi người ta.
1.1 hw1: trăm năm
1.1 nt1: Thời gian một đời người ta ở trần ...

1.2 img:/data/kieu/1866/page01a.jpg
1.2 hn: 𡨦才𡨦命窖𠳋恄饒
1.2 qn: Chữ tài chữ mệnh khéo là ghét nhau.
...
```

```
<text>
  <body>
    <ab xml:id="1-2">
      𥙿𪗋𪚥揆𠊛些<lb/>
      𡨦才𡨦命窖𠳋恄饒<lb/>
      Trăm năm<note place="inline" n="1">trăm năm: Thời gian một ...</note>
      trong cõi người ta.<lb/>
      Chữ tài chữ mệnh khéo là ghét nhau.
    </ab>
...
```

Fig. 2:
*Left: the original plain text format in Nôm script (*hn*) and Quốc Ngữ (*qn*) including the commentary (*hw1 *and* nt1*) for the first two lines of Truyện Kiều (1866). Right: excerpt of the data transformed into a XML-format readable by LERA, where the two lines have been merged into one segment. The philological commentary was transformed to a <note>-tag.*

Collation with LERA

In order to display Nôm script properly, an appropriate font [3] was embedded into LERA. To make the text segments more legible, options for displaying the inserted line breaks and centering the text have been added (see Fig. 3).

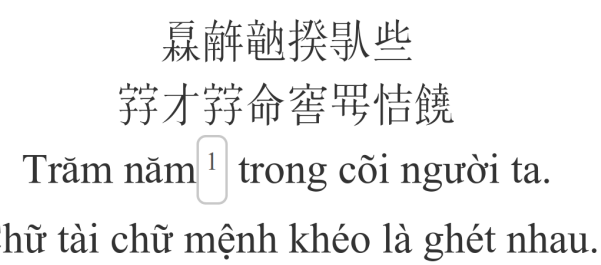


Fig. 3:
Representation of a text segment of Truyện Kiều in LERAs full text visualization. Philological commentaries are indicated by consecutive numbers, with their text fading in via mouse-over.

LERA uses a two-step approach for collation (Pöckelmann et. al 2022), both applied to the four-line segments: an alignment of text segments according to their similarity and a detailed comparison of the aligned segments.

For the first step the manual assignment already encoded via line numbers into the data can be used for Truyện Kiều. However, the algorithm for alignment integrated in LERA produces nearly the same alignment.

The detailed comparison reveals many textual variants spread over the entire length of the work for both scripts. This is still true for modern Vietnamese when minor variants, like differences in capitalization and punctuation marks, or even diacritics, which are very important for a word's meaning, are ignored with the aid of LERAs filter system. Overall, textual variants occur more frequently in the Nôm script versions than in the modern Vietnamese versions. The reasons for this are almost all in the complex phonetic-semantic-composition of characters. Phono-semantic compound characters (such as 坦) have two components: one suggesting the word's meaning (土) and the other its approximate sound (旦). The characters 旦 and 坦 (used in the last witness only) have the same meaning *đắn*. There are also rules in the positioning depending on function: semantic components tend to appear on the top or left side of characters; phonetic components at their bottom or right side, e.g.: 撑 = ⿰扌 掌. Variant characters, which are homophones and synonyms caused by allography, occur also in the work, e.g., 句 and 勾, which were both translated to *câu,* vary in their phonetic component because the second uses with 厶 another written representation of 口. The Nôm script characters have been compared so far only as a whole, so those variants could not be further differentiated.

It should be noted that the fixed structure of the segments with the verse form makes the comparison somewhat easier than for other texts: with a few exceptions, there are only substitutions of characters and no insertions or deletions.

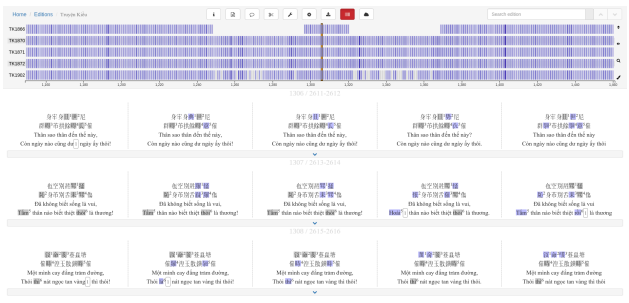


Fig. 4:
Screenshot of LERA showing an excerpt of the five collated version of Truyện Kiều. The navigation bar shows a very similar structure for all witnesses (with the exception of 1866, recognizable by the gaps at the top), while the density of the blue highlightings indicate textual variance within all aligned segments (the darker the more different the aligned segments are).

Discussion and Future Work

Only minor adaptations of LERA were necessary to allow a basic comparison of the text witnesses of Truyện Kiều, such as the integration of the font or a slightly adapted display of text segments.

The work shows further potential for the improvement of LERA for an application to Asian languages. This includes an approach to process, normalize and compare the components of complex composite characters separately and represent the results in an appropriate way. Furthermore, the demand for integrating the top-to-bottom writing direction [4] became apparent.

Bibliography

Alder, E. and Cranfield, J., Dryden, L., Duncan, I., Ferguson, C., James, S.J., Kerr, D., Luckhurst, R., Machin, J., Schwan, A. and Wild, J. (2020). *Edinburgh Conan Doyle Project*. https://edinburgh-conan-doyle . org/

Balaban, J., Collins, L., Lesser, S., Phan, J., Schmid, D. N. and Việt, N. T. (2001). *Vietnamese Nôm Preservation Foundation.* http://www.nomfoundation.org/

Bragagnolo, M. (2017). *HyperAzpilcueta – Visualizing the instability of early modern normative knowledge*. https://www.rg.mpg.de/research-project/hyperazpilcueta

Bremer, T. (2013). *Guillaume Thomas Francois Raynal: Histoire philosophique et politique des établissements et du commerce des européens dans les deux Indes. Semi-automatische Di ff erenzanalyse von komplexen Textvarianten*. https://www.izea.uni-halle.de/forschung/d-erschliessungsprojekte-und-editionen/guillaume-thomas-francois-raynal.html

Gründler, B., Pöckelmann, M. (2018). *Adjusting LERA for the comparison of arabic manuscripts of kalıla wa-dimna.* In: Digital Humanities 2018 - Book of Abstracts. pp. 467–468. Mexico City, Mexico (2018)

Gründler, B. (2019). *Kalıla and Dimna – AnonymClassic.* https://www.geschkult.fu-berlin.de/en/e/kalila-wa-dimna/

Gründler, B., van Ginkel, J.J., Redwan, R., Khalfallah, K., Toral, I., Stephan, J., Keegan, M.L., Beers, T.S., Kozae, M., Ahmed, M.M. (2020). *An interim report on the editorial and analytical work of the AnonymClassic project*. Medieval Worlds (11), 241–279

(2020), https://doi.org/10.1553/medievalworlds_no1 1_2020s241

Hahn, B., Eusterschulte, A., Kieslich, I. and Pischel, C. (2020). *Hannah Arendt Digital – Kritische Gesamtausgabe.* https://hannah-arendt-edition.net

Molitor, P., Necker, G., Pöckelmann, M., Rebiger, B., Ritter, J. (2020). *Keter Shem Tov – Prozessualisierung eines Editionsprojekts mit 100 Textzeugen*. Conference of the Digital Humanities Association in German-speaking Countries (DHd). Paderborn (2020)

Necker, G., Molitor, P., Ritter, J., Rebiger, B., Pöckelmann, M. (2019). *Kabbalah Editions.* https://kabbalaheditions.org/

Pöckelmann, M., Medek, A., Ritter, J. and Molitor, P.(2022). *LERA - An interactive platform for synoptical representations of multiple text witnesses*. In: Digital Scholarship in the Humanities (DSH). Oxford University Press 2022, https://doi.org/10.1093/llc/fqac021

Notes

1. Homepage of LERA – Locate, Explore, Retrace and Apprehend complex text variants: https://lera.uzi.uni-halle.de/
2. See: https://tei-c.org/
3. The font *Nom Na Tong* is also provided by the Vietnamese Nôm Preservation Foundation at https://github.com/nomfoundation/font/releases/tag/v4.8
4. Right-to-left has been added in the course of scholarly editions of Arabic and Hebrew texts.

A 3-D analytic framework of humanistic objects: data modeling paradigms, computational analysis and close reading

Maeir, Noam
noam.maeir@mail.huji.ac.il
Hebrew University, Israel

Keydar, Renana
renana.keydar@mail.huji.ac.il
Hebrew University, Israel

Introduction

Recently, distant and close reading methods, broadly interpreted as quantitative-computational and qualitative analyses, have emerged as the primary methods of scholarship in DH (Schöch et al, 2020). Consequently, scholars have been developing various conceptual models attempting to combine the two methods in a systematic and comprehensive manner (Keydar 2019, 2020; Roe et al, 2020). In addition, aside from conceptualizing the relationship between distant and close reading, the "datafication" of humanistic data has brought scholars to emphasize the multidimensionality of their data (Windhager 2020). That is, humanistic data is not comprehensively represented via any single modeling paradigm. Therefore, researchers are calling for analytic frameworks that apply multiple modeling paradigms to the study of data.

Building on these recent developments, in this paper we present a new conceptual framework for the analysis of humanistic objects, whereby close and distant reading are applied within multiple data modeling paradigms. In other words, we suggest that the study of humanistic objects is to be conducted within 3 dimensions – representational (data modeling); computational (distant reading); qualitative (close reading).

The significance of the 3D framework, and especially the inclusion of the data modeling aspect, will be demonstrated through a case study from the field of comparative religion and the study of late antiquity.

Case study

Our case study deals with the cultural interactions between Greek and Syriac literary cultures during the first millenium CE (2 nd – 10 th centuries CE). Traditionally, the interaction between the two has been characterized by scholars as expressing a gradual "Hellenization" of Syriac culture, i.e., an increasing influence of Greek upon Syriac culture (e.g.: Brock 2012; 2015; Butts 2016). The recent massive digitization of manuscripts and texts offers an opportunity to re-examine the paradigm of Syriac Hellenization, within our 3D conceptual framework.

Representational analysis

A survey of the available corpora of Syriac literary artifacts results in the identification of two kinds of data found in online databases – texts (.txt files) and manuscript images (.jpg files) – each representing a different data modeling paradigm of Syriac literature. In other words, the results of the analysis of the representational dimension of

Syriac literary culture include two data representations, or modeling paradigms, one that enables to study texts, and the other that enables to study the images of manuscripts. Therefore, our distant and close reading of Syriac-Greek interactions will be conducted within each data modeling paradigm.

Literary culture as texts

Operating within a modeling paradigm that represents Syriac literature as texts, the emerging cultural agents – those that produce Syriac literary culture – are Syriac authors, and the analysis, computational and qualitative, of authored texts is the manner by which cultural interactions are discovered.

Consequently, in line with previous scholarship, our distant and close reading of Syriac texts resulted in the identification of a gradual process of Syriac Hellenization, as seen, for instance, in the increased appearance of Greek loanwords and adjectival forms. Nevertheless, as will be seen, shifting to the second modeling paradigm will yield different results.

Literary culture as manuscript images

When Syriac literary culture is represented as manuscript images, the Syriac cultural agents that become apparent are the scribes of the manuscripts – those that produced the manuscripts in their entirety – as well as the authors of the texts that appear in them. Yet, as many of the Syriac manuscripts present compilations of many different authored texts – i.e., Multiple Text Manuscripts (Kessel 2015; Fiori 2020) – the importance of the scribe is highlighted. As such, our hybrid distant-close reading will be centered upon scribal actions (rubrications, omissions, erasures, notations, etc.).

As opposed to the previous section's analysis, the study of scribal actions reveals an increasing presence of scribal texts (i.e., longer titles and notes) as well as a decrease in the presentation of authored texts (i.e., shorter excerpts of longer texts). In other words, the analysis of data from a modeling paradigm that represents Syriac literature as manuscript images, resulted in the identification of an opposed cultural process to the traditional Hellenization paradigm. That is, from the perspective of this modeling paradigm, Syriac literary culture gradually distances itself from the earlier Greek forms to add distinct Syriac features – not Hellenization, but Syriacization. Accordingly, Syriac-Greek cultural interactions in late antiquity are characterized by a decrease of Greek influence upon Syriac culture.

Conclusion

To conclude, this paper will present a new 3D conceptual framework for the analysis of humanistic objects, that integrates multiple data modeling paradigms with a hybrid analysis of close and distant reading. The results of our case study demonstrate that a given modeling paradigm governs and confines the applications of distant (computational) and close (qualitative) readings, to a limited scope of analysis, which can be contradicted via a different modeling paradigm. Therefore, the need to include multiple modeling paradigms within the study of humanistic data appears to be crucial. Furthermore, future scholarship should theorize the relationship between the different modeling paradigms and their derived, and in our case opposing, results.

Bibliography

Butts, A. M. (2016). *Language Change in the Wake of Empire: Syriac in its Greco-Roman Context*. Indiana: Eisenbrauns.

Brock, S. P. (2012). A tentative check list of dated Syriac manuscripts up to 1300.*Hugoye*, 15: 21-48.

Brock, S. P. (2015). Charting the Hellenization of a literary culture: the case of Syriac.*Intellectual History of the Islamicate World*, 3: 98–124.

Fiori, E. (2020).*Florilegia Syriaca.* Mapping a knowledge-organizing practice in the Syriac world, Ca' Foscari University of Venice, 30 January–1 February 2020. *COMSt Bulletin* 6(1): 93-109.

Kessel, G. (2015). Syriac monastic miscellanies.*COMSt,* 1: 411-414.

Keydar , R. (forthcoming). Changing the lens on survivor testimony: topic modeling the Eichmann trial . *Jewish Studies Quarterly*.

Keydar, R. (2020). Listening from afar: an algorithmic analysis of testimonies from the international criminal courts. *Illinois Journal of Law, Technology & Policy,* 1: 55-83.

Keydar, R., Litmanovitz, Y., Hasisi, B., and Kantor, Y. (forthcoming). Modeling repressive policing: computational analysis of protocols from the Israeli state commission of inquiry into the October 2000 events. *Law and Social Inquiry*.

Roe, G., Gladstone, C., and Olsen, M. (2020). Mind the gap: bridging distant and close reading across heterogeneous text collections. *ADHO*.

Schöch, C., Eder, M., Arias, R., Francois, P. and Primorac, A. (2020). Foundations of distant reading: historical roots, conceptual development and theoretical

assumptions around computational approaches to literary texts. *ADHO*.

Wenger, M., Kalir, T., Keydar, R., Stanovsky, G. (2021). A *utomated extraction of sentencing decisions from court cases in the Hebrew language. Proceedings of the 2021 Conference on Natural Legal Language Processing* (NLLP 2021).

Windhager, F., Salisu, S. and Mayr, E. (2020). Reassembling elephants: a multi-spatiotemporal visualization method for history and humanities data. *ADHO*.

Towards Adaptive Retrieval Systems for Intertextuality Research: a case study on Biblical Intertextuality

Manjavacas Arevalo, Enrique
enrique.manjavacas@gmail.com
University of Leiden, Netherlands

In cultural studies, intertextual theory is concerned with the links that literary texts establish with each other through different types of referencing. Through the recognition and interpretation of these links, new readings and perspectives on literary works open up, but the process of manually identifying the underlying intertextual networks in a sufficiently exhaustive manner is an arduous and laborious process in which literary scholars are left to their own erudition. Increasingly, computational methods are applied in this domain with the goal of automating the retrieval of intertexts, which range from more or less exact quotations, to subtler allusions or paraphrases.

For a given research project, the choice about the retrieval algorithm is typically done on the basis of the type and style of intertext that are expected. For example, detecting near-verbatim textual borrowings spanning long sequences across 19th century newspapers (D. A. Smith, Cordell, and Dillon 2013; D. A. Smith et al. 2014) is best served with a text alignment algorithm (e.g. the Smith-Waterman algorithm for local alignment: T. F. Smith and Waterman 1981). In contrast, when tackling non-literal intertextuality (e.g. literary allusions), lexical semantics needs to be taken into account by using lexical databases like WordNet (Fellbaum 2012) or deploying distributional semantic models in order to capture different types of semantic relatedness like synonymy or antonymy in an unsupervised manner (Bamman and Crane 2008; Manjavacas, Long, and Kestemont 2019; Moritz et al. 2016).

However, in a common research scenario, characterized by the task of extracting intertextual connection from a given target collection to a second collection serving as the source, it is not always clear what type and style of intertextuality may be found, and, even if it were, common retrieval algorithms include a variety of parameters that have to be tweaked manually in order to achieve good retrieval performance (Büchler et al. 2014). For these reasons, a long term desideratum for computational research in intertextuality is to develop general methods that can adapt to the style of intertextuality present in the given textual collections (Manjavacas Arévalo 2021) **.** Currently, the main hurdle to develop this type of approaches is the lack of benchmark datasets and, ultimately, the inherent difficulty to produce them (Manjavacas Arevalo, Mellerin, and Kestemont 2021).

In the present work, we approach the topic of automatically adapting text reuse algorithms to the target corpus from a different angle. We focus on the Smith-Waterman algorithm for local text alignment, introducing an extension module that exploits target corpus information in order to boost the retrieval performance. The Smith-Waterman algorithm computes an alignment score for two given input texts that corresponds to the optimum sequence alignment. This optimum alignment score depends on the relative importance that is given to the match, mismatch and gap operations with which the alignment is computed. These relative importances are free parameters that ultimately control the type of intertext that is matched (e.g. a lower gap penalty would license long textual borrowings even if other text has been intercalated in between).

In this work, we focus on the match score. This score is added to the output alignment score for every word in the target sequence that matches a word in the source sequence (e.g. words that share the same underlying lemma). While traditional implementations of alignment algorithms for intertextuality treat this score in a monolithic way, keeping it constant regardless of the particular words that are being matched, we suspect that generating word-specific match scores can be a powerful yet uncomplicated approach in order to adapt the retrieval algorithm to the particular type of intertext present in a given collection. For example, if an author is more likely to place an intertextual connection when discussing a particular topic, matches on words related to that topic could be boosted. Moreover, such topic-related trends could be estimated from corpora annotated with intertextuality.

In our contribution, we focus on frequency-based trends and test an approach that aims at boosting matches on words that carry more information. Following the same intuition that underlies the common frequency-based weighted scheme for document representation known as Tf-IDF, we generate word-specific match scores on the basis of how specific the words are for the underlying

collections. While being adaptive to the underlying corpus, this approach doesn't require pre-existing annotations regarding intertextuality and still shows significant retrieval gains.

We benchmark the proposed extension against the standard Smith-Waterman algorithm, as well as against two representative algorithms belonging to the Vector Space Model family. Moreover, in contrast to common practice in computational intertextuality, we put the proposed extension to test across a variety of benchmark datasets as well as three different languages (Latin, Ancient Greek and Early Modern English). The experiments show that the proposed scoring system based on word frequencies and frequency ranks outperforms the alternative approaches in almost all benchmark datasets, while only adding two extra tunable hyper-parameters to the Smith-Waterman algorithm.

In this talk, we would like to present the approach and the evaluation in detail, as well as showcase in what particular ways the proposed method boosts retrieval performance by means of an exhaustive error analysis. By making the code and datasets available, we look forward to helping other researchers in the field with the automatic retrieval of intertext as well as with the systematic evaluation of future retrieval approaches.

Bibliography

Bamman, David, and Gregory Crane. 2008. "The Logic and Discovery of Textual Allusion." In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data*. http://www.perseus.tufts.edu/{~}ababeu/latech2008.pdf.

Büchler, Marco, Philip R. Burns, Martin Müller, Emily Franzini, and Greta Franzini. 2014. "Towards a Historical Text Re-Use Detection." In *Text {Mining}*, 221–38. Springer. https://doi.org/10.1007/978-3-319-12655-5_11.

Fellbaum, Christiane. 2012. "WordNet." In *The Encyclopedia of Applied Linguistics*. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781405198431.wbeal1285.

Manjavacas Arévalo, Enrique. 2021. *Computational Approaches to Intertextuality. from Retrieval Engines to Statistical Analysis : Thesis.* Proefschriften UA-LW : Letterkunde: 2021: 2. https://doc.anet.be/docman/docman.phtml?file=.irua.6b6d46.178931.pdf.

Manjavacas Arevalo, Enrique, Laurence Mellerin, and Mike Kestemont. 2021. "Quantifying Contextual Aspects of Inter-Annotator Agreement in Intertextuality Research." In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 31–42. Punta Cana, Dominican Republic (online): Association for Computational Linguistics. https://aclanthology.org/2021.latechclfl-1.4.

Manjavacas, Enrique, Brian Long, and Mike Kestemont. 2019. "On the Feasibility of Automated Detection of Allusive Text Reuse." In *Proceedings of the 3rd Joint {SIGHUM} Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 104–14. Minneapolis, USA: Association for Computational Linguistics. https://www.aclweb.org/anthology/W19-2514.

Moritz, Maria, Andreas Wiederhold, Barbara Pavlek, Yuri Bizzoni, and Marco Büchler. 2016. "Non-Literal Text Reuse in Historical Texts: An Approach to Identify Reuse Transformations and Its Application to Bible Reuse." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1849–59. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1190.

Smith, David A., Ryan Cordel, Elizabeth Maddock Dillon, Nick Stramp, and John Wilkerson. 2014. "Detecting and Modeling Local Text Reuse." In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 183–92. https://doi.org/10.1109/JCDL.2014.6970166.

Smith, David A., Ryan Cordell, and Elizabeth Maddock Dillon. 2013. "Infectious Texts: Modeling Text Reuse in Nineteenth-Century Newspapers." In *2013 IEEE International Conference on Big Data*, 86–94. IEEE. https://doi.org/10.1109/BigData.2013.6691675.

Smith, T.F., and M.S. Waterman. 1981. "Identification of Common Molecular Subsequences." *Journal of Molecular Biology* 147 (1): 195–97. https://doi.org/10.1016/0022-2836(81)90087-5.

Aplicación de las Humanidades Digitales a los estudios sobre Teatro de la Antigüedad Clásica

Una revision de las dos últimas décadas (2001-2020)

Martínez Nieto, Roxana Beatriz
roxana.martinez@unir.net
Universidad Internacional de la Rioja UNIR, Spain

A partir del creciente interés suscitado por humanistas y estudiosos del mundo clásico en la investigación y tratamiento digital de datos sobre la Antigüedad clásica, y más en concreto sobre el teatro y la transmisión de los trágicos griegos y latinos, se aborda en el presente

en el presente artículo una presentación de los estudios más relevantes realizados durante las últimas dos décadas (2001-2020). Es innegable que el campo de las Humanidades Digitales aplicado a la investigación textual de nuestros clásicos antiguos está abriendo nuevas líneas de actuación que acercan la investigación a la enseñanza y ofrecen nuevas perspectivas en el campo de la enseñanza en entornos virtuales sobre el teatro de época clásica. El presente artículo evidencia finalmente el gran avance en el campo del análisis computacional puesto al servicio de la interpretación textual tradicional y su repercusión en performance y tratamiento estructural de las piezas trágicas más representativas de la Antigüedad clásica.

Propuesta de investigación: Partiendo de la premisa de que las humanidades digitales son un campo de investigación al servicio de los estudiosos del mundo antiguo, filólogos, estudiosos del teatro clásico e historiadores de la Antigüedad clásica, se ofrece una aproximación a la interpretación del teatro clásico a partir de los estudios realizados por humanistas digitales en las últimas dos décadas. Las ediciones digitales y los estudios de crítica textual, junto al análisis computacional del léxico teatral de época clásica permiten actualmente obtener resultados cuantitativos que a su vez redundarán en resultados cualitativos sobre la enseñanza en entornos virtuales y puesta en escena del teatro clásico.

Marco metodológico: La entrada de las tecnologías digitales en el mundo de las humanidades y del teatro es una realidad que confirma la nueva dimensión cultural de la sociedad actual. Los modelos educativos en entornos virtuales también ponen de manifiesto la transformación en la enseñanza del siglo XXI, se hacen eco de los nuevos retos docentes y ofrecen innovadoras perspectivas en aplicaciones digitales al ámbito académico. Luego de entender el contexto de crecimiento dinámico social en el que se insertan las Humanidades Digitales como disciplina, cabe citar la constitución de Proyectos de investigación, formación de Asociaciones, grupos específicos de investigación en un campo determinado, la celebración de Congresos *ad hoc* y la proliferación de revistas sobre el Humanidades Digitales de carácter internacional. Los materiales publicados en el ámbito de estudio teatral de la Antigüedad clásica conforman el propósito de análisis del presente trabajo y la metodología de investigación establecida se basa en la identificación de los estudios relevantes, temática de su contenido y transferencia de la información a otros ámbitos de estudio, a partir de su aplicación y divulgación.

Discusión de naturaleza crítica: La interpretación del drama griego y latino a la luz de las Humanidades Digitales plantea una discusión de naturaleza crítica, la cual pretende arrojar luz sobre el *status quaestionis* en torno al tratamiento digital de textos teatrales clásicos. El acercamiento al estudio del teatro clásico a través de aplicaciones como Shiny DraCor, la marcación en lenguaje XML-TEI y el análisis cuantitativo del léxico puede dar lugar a nuevas interpretaciones filológicas acerca de los patrones del teatro clásico, su estructura interna y su proyección hacia el público. Concretamente a partir de los resultados cuantitativos disponibles actualmente, se persigue extraer unos hallazgos cualitativos que nos permitan interpretar los modelos de performance teatrales objeto de estudio, mediante el cotejo de patrones tradicionales de la tragedia antigua, que a su vez incluye el análisis de la presencia tanto de personajes principales y secundarios, como de intervenciones por parte del coro dentro de una misma pieza dramática. La interrelación resultante puede ser confirmada o desestimada por la evidencia de los datos digitales extraídos, ayudando sobremanera al estudio de las redes sociales implícitas en el teatro clásico antiguo. Todo el proceso será tratado partiendo de las obras conservadas de los trágicos griegos Esquilo, Sófocles y Eurípides y podrá extenderse al análisis estructural de otras piezas teatrales clásicas. Se persigue asimismo una aplicación real de estos resultados en el aula considerando aspectos pedagógicos y académicos que ayuden al docente a trasladar los resultados de investigación en HD sobre teatro clásico antiguo. En cuanto a la relación entre investigación y puesta en escena, serán los datos digitales sobre reconstrucción en 3D los que ayuden a ofrecer una visión de los espacios teatrales en entornos virtuales no disponible hasta el momento. Ello hará posible describir nuevas herramientas para el análisis textual digital aplicadas a los textos teatrales clásicos, que podrían producir importantes ideas disciplinarias y trasladar de forma efectiva las actuales investigaciones en HD al aula dentro de un futuro muy cercano.

Valor de la contribución teórica a las humanidades digitales: Los resultados obtenidos ayudan a demostrar que actualmente los humanistas digitales y estudiosos de la época clásica, quienes se dedican a investigar sobre teatro de origen griego y latino, centran sus investigaciones principalmente en el análisis digital cuantitativo de datos, con el fin de llegar a comprender las relaciones dialógicas entre personajes centrales dentro de una misma obra o entre piezas distintas. Se hace necesario, por ende, profundizar en el sentido propedéutico del proceso de enseñanza-aprendizaje de los distintos modelos de teatro, que se están ofreciendo en las aulas actuales, cada vez más integradas en entornos virtuales; y se consideran campos abiertos a la exploración y la investigación los análisis cuantitativos y estadísticos, el tratamiento digital de textos y el análisis de la estructura de los patrones teatrales, entre otras vías de investigación. Cabe señalar además la necesidad de seguir indagando en el campo de las realidades virtuales, especialmente en las réplicas de espacios teatrales en el mundo antiguo, reconstrucciones en 3D de artefactos antiguos y patrones artísticos, considerados

nuevas áreas de investigación y futuras líneas de desarrollo y prospectiva. Los avances tecnológicos, así como su directa implementación en la investigación dedicada al drama clásico, resultan enormemente fructíferos para los estudios académicos en torno a problemas de interpretación, representación escénica y crítica textual tradicionales, que podrán ser reconsiderados desde la nueva perspectiva que ofrecen los tratamientos digitales en del campo de las Humanidades, contribuyendo a acercar al gran público el teatro clásico.

Bibliografía

Bozia, E. (2018). Reviving Classical Drama: virtual reality and experiential learning in a traditional classroom. *Digital Humanities Quarterly,* 12(3): 1-19.

Laan, N. M. (1995). Stylometry and Method: the Case of Euripides. *Literary and Linguistic Computing.* 10(4): 271-278.

Slaney, H. (2017). Motion Sensors: Perceiving Movement in Roman Pantomime. In Betts, E. (ed.) *Senses of the Empire: Multisensory Approaches to Roman Culture.* Routledge. New York and London (2017): 159-175

Vervain, C. Performing ancient drama in mask: The case of Greek new comedy. *New Theatre Quarterly*. 20(3), pp. 245-264.

From Modern to Medieval: Detecting and Visualizing Entities in Manuscripts of Marco Polo's Devisement du Monde

Meinecke, Christofer
cmeinecke@informatik.uni-leipzig.de
Leipzig University, Germany

Wrisley, David Joseph
djw12@nyu.edu
New York University Abu Dhabi

Jänicke, Stefan
stjaenicke@imada.sdu.dk
University of Southern Denmark

INTRODUCTION

Digital humanities have had a strong focus on algorithmic reading of textual data, which has been brought together with visualization for comparing and analyzing results (Jänicke et al., 2017). In recent years, this textual focus has grown to include other modalities such as audio-visual data (Arnold and Tiltion, 2019; Wevers and Smits, 2020). In turn, methods in computer vision (Arnold and Tiltion, 2019) have been proposed for the specificities of audio-visual corpora. As a starting point for distant viewing of medieval illumination, we applied computer vision methods to a dataset of images from manuscripts of the French Marco Polo textual tradition, images that demonstrate a strong visual coherency. Extant in some 15 manuscripts, the Devisement du monde is famous for descriptions of extra-European travel and the depiction of the exotic wonders of Asian cities (Cruse, 2019). We set out to see if repeated visual features across this image corpus are detectable using object detection, what visualization would allow us both to understand better Polo's depiction of the exotic and how modern image hierarchies might be adapted to the specificities of medieval manuscripts.

For image classification and object detection of images, large datasets with class hierarchies exist like ImageNet (Deng et al., 2009) and Open Images (Kuznetsova et al., 2020). These datasets and their underlying hierarchies are neither particularly effective at identifying the wide variety of entities depicted in medieval manuscripts, nor do they detect entities well given the representational density of medieval illumination. Furthermore, there is insufficient data for training these neural networks. In our work, we argue that networks trained on natural image datasets can provide both a first impression (Crowley and Zisserman, 2014), and a convenient starting point for building new classes and hierarchies and they can be even used to extract some initial training samples from small- to medium-sized image corpora.

We applied computer vision methods on a dataset of some 700 medieval illuminations from seven manuscripts and built a visual interface to explore and annotate the results. We were interested in the possibility of editing the classes of contemporary hierarchies, replacing them with categories more appropriate for the period and the corpus.

DATA & IMAGE PROCESSING

Each image shows a page with a visual scene depicting different aspects of Polo's description. We applied object detection by using the Faster R-CNN (Ren et al., 2015) trained on Open Images. The label hierarchy of Open Images consists of 600 different classes including parent and

child relations. The object detection extracts 100 bounding boxes for each image with a confidence score and a label for the detected entity. The result was 71,400 bounding boxes. Furthermore, we extracted image embeddings for each bounding box detected with an EfficientNet (Tan and Le, 2019) trained on ImageNet. For the image embeddings, we use faiss (Johnson et al., 2019) to query the most similar bounding boxes for each example based on the Euclidean distance between the embeddings. This allows us to see the most similar parts of another image to an image of interest.

VISUAL INTERFACE

The design of the visual interface facilitates exploration of the image dataset and comparison with different depictions of specific entities. For this, the object classes can be accessed through a Tag Cloud where frequency is encoded by font size, or through a tree that visualizes the Open Images hierarchy with all classes found in the Marco Polo dataset. Such interfaces for visual exploration and annotation allow the professional viewer/reader to focus on a given interest to annotate new areas or investigate the objects found inside the image, delete them or even edit their labels (Siemens et al., 2009). To prevent visual clutter, they can filter by confidence value and select or deselect specific classes. When focusing on one specific bounding box, it is also possible to display the bounding boxes that intersect, that are inside or outside the box of interest. Figure 1 shows a page of the dataset with the entities found by the neural network. For a given object class, all depictions are displayed in a 2D grid ordered by the confidence score assigned by the neural network. Examples can be seen in Figures 2 and 3. The interface is designed for both discovery and revision by clicking on a bounding box of interest, which leads us to see the most similar bounding boxes. It is also possible to select multiple bounding boxes and delete or re-label them in case of an imprecise classification. Furthermore, the expert viewer is able to annotate areas in the image with new classes, thereby contributing to a new category in the Tag Cloud (Annotated) and transforming it into a TagPie (Jänicke et al., 2018), seen in Figure 4.

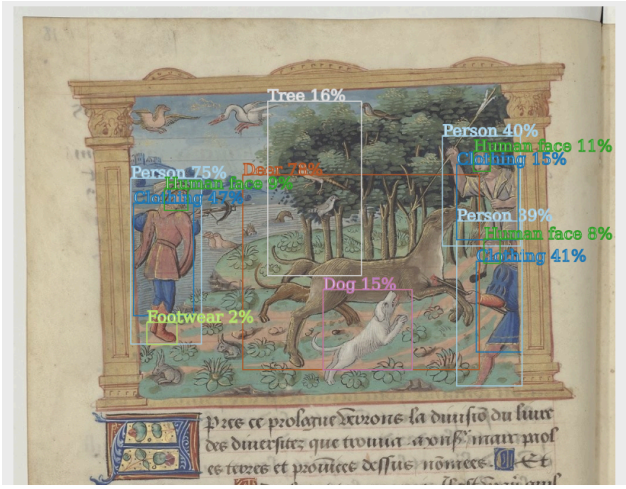


Figure 1:
A page of the dataset with entities found by a neural network in an illumination from BnF Arsenal ms 5219



Figure 2:
An overview of samples of faces marked by the highest confidence scores



Figure 3:
An overview of samples of human figures found in the corpus with the highest confidence scores



Figure 4:
The TagPie gives an overview of the classes found by the neural network (green) and those from human annotations (purple)

DISCUSSION

Whereas some anachronistic categories were persistent throughout the output of the initial system, other objects such as those mentioned in Figures 1 - 3 led to rather convincing recognition. Furthermore, the summary views of the visual interface proved particularly effective at demonstrating the tensions found between the codified visual languages of medieval French manuscripts and the diachronic innovative attempts at representing the "unprecedented images of the world beyond Europe's borders" as well as domains in which patterns in those tensions were particularly pronounced (Keene, 2019: 196).

On the other hand, the interface created to explore, revise and manipulate features in the Marco Polo visual corpus provides us with a stepping stone for working with larger visual corpora built from across the global middle ages. As our inquiry evolves, finding ways to guide the viewer from the extracted objects and their computed confidence levels back to full images and relevant metadata will be crucial for allowing for sufficient contextualization to facilitate interpretation. Furthermore, our current method for revision and addition of labels is open-ended, but in future work, we intend to lead the annotation toward established art historical vocabularies to ensure future discoverability. Future work will also focus on ways to achieve the "best of both worlds," allowing research to move from the modern to the medieval, that is, for current day hierarchies to be adjusted and augmented by domain- and period-specific terminology with the support of expert knowledge.

Creating this visual pathway for visual exploration and hypothesis generation using computer vision techniques is not a trivial task, since the metadata of legacy databases of manuscript illumination (Mandragore, Initiales, Digital Scriptorium, etc.) also vary in both size and granularity. Furthermore, there is a need for methodologies to combine or unify the vocabulary of different datasets, bridge the gap between general and domain-specific vocabularies, as well as to create expert hierarchies of entities found in manuscript illumination in order to create appropriate training datasets to deal with issues of cross-depiction (Hall et al., 2015).

Bibliography

Jänicke, S. Franzini, G. Cheema, M.F. and Scheuermann, G . (2017). TagPies: Visual text analysis in digital humanities. Computer Graphics Forum 36, 6 2017, 226–250.

Arnold, T. and Tilton, L. (2019). Distant viewing: analyzing large visual corpora. Digital Scholarship in the Humanities 34, Supplement_1, i3–i16.

Wevers, M. and Smits, T . (2020). The visual digital turn: Using neural networks to study historical images. Digital Scholarship in the Humanities 35, 1 (2020), 194–207

Cruse, M .(2019). Novelty and Diversity in Illustrations of Marco Polo's Description of the World.

Keene, B.C .(2019).*Toward a Global Middle Ages: Encountering the World Through Illuminated Manuscripts.* Los Angeles : J. Paul Getty Museum.

Deng, J. Dong, W. Socher, R. Li, L.J. Li, K. and Fei-Fei, L . (2009). ImageNet: A Large-Scale Hierarchical Image Database. In CVPR09.

Kuznetsova, A. Rom, H. Alldrin, N. Uijlings, J. Krasin, I. Pont-Tuset, J. Kamali, S. Popov, S. Malloci, M. Kolesnikov, A. Duerig, T. and Ferrari, V. (2020). The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision 128, 7 (2020), 1956—1981.

Crowley, E.J. and Zisserman, A . (2014). In search of art. In European Conference on Computer Vision. Springer, 54–70.

Ren, S. He, K. Girshick, R. and Sun, J . (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems. 91–99.

Tan, M. and Le, Q . (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. International Conference on Machine Learning 6105–6114

J ohnson, J. Douze, M. and Jégou, H . (2019). Billion-scale similarity search with gpus. IEEE Transactions on Big Data

Siemens, R. Leitch, C. Blake, A. Armstrong, K. and Willinsky, J. (2009). "It May Change My Understanding of the Field": Understanding Reading Tools for Scholars and Professional Readers. Digital Humanities Quarterly, 3, 4.

Jänicke, S. Blumenstein, J. Rücker, M. Zeckzer, D. and Scheuermann, G . (2018). TagPies: Comparative Visualization of Textual Data. VISIGRAPP (3: IVAPP) 40–51.

Hall, P. Cai, H. Wu, Q. and Corradi, T . (2015). Cross-depiction problem: Recognition and synthesis of photographs and artwork. Computational Visual Media 1, 2, 91–103.

A Five-Star Model for Linked Humanities Data Usability

Middle, Sarah
sarah.middle@open.ac.uk
Open University, National Museums Scotland

Introduction

Of the various approaches to modelling Humanities data, Linked Data is particularly effective for representing complexity and nuance, while facilitating implementation of the FAIR data principles (Findable, Accessible, Interoperable, Reusable) (Wilkinson et al., 2016). However, despite a recent increase in Linked Humanities Data production, uptake remains low, partly due to usability issues. This paper presents key findings from a study that investigated these issues by focusing on the Ancient World, a subject domain where Linked Data implementation is relatively mature and which, due to its multi-faceted nature, could be considered a microcosm of Humanities research. My study sought to establish user and producer needs through a survey of the research community and detailed interviews with selected participants. Responses informed a series of recommendations, presented as a *Five-Star Model* for Linked Humanities Data usability.

Background

The term 'Linked Data' refers to a set of technologies to describe entities, such as places, people, or objects, and connect them based on features they have in common. A Linked Data approach can provide significant benefits for Humanities research: its rich semantic descriptions, disambiguation capabilities, and interoperability can unlock opportunities to address new research questions and reveal previously undiscovered relationships. However, previous research has identified barriers to producing Linked Humanities Data, including unfamiliarity with a graph data model (Barbera, 2013: 96; Ross et al., 2015: 118), the significant time investment required for training, and difficulties in securing support (Isaksen, 2011: 153–54; Geser, 2016: 12, 56; Granados-García, 2020: 261–64). Where time is short and training/support are limited, Linked Data producers naturally prioritise immediate project goals over long-term usability.

Recommendations to assist the production of usable Linked Humanities Data tools and resources are therefore required. Progress has been made by the 'Linked Open Usable Data' (LOUD) initiative, resulting in the definition of five 'Design Principles'(Linked Art Contributors, no date); however, their primary aim is usability by developers, rather than end users. Elsewhere, user consultations have been conducted in relation to specific tools and resources, such as *Europeana* (Angelis et al., 2015) and *Recogito* (Simon et al., 2015). My study differed from these in investigating the usability of multiple tools and resources across the wider digital ecosystem.

Methodology

The first phase of my study comprised a survey of Ancient World researchers. Assuming Linked Data use among this audience would be relatively low and intending to gain broader insights, I aimed the survey at all Ancient World researchers who use digital tools or resources, with specific questions about Linked Data use and production. The survey ran during spring 2018 and received 212 responses. The second phase involved interviews with 16 survey participants to explore their experiences in more depth (between autumn 2018 and spring 2019 [1]). Interview participants had differing levels of technical experience, while being broadly representative of the survey demographics.

Findings

From analysing participant responses, I found that usability can be facilitated by considering six key factors when planning a Linked Humanities Data project: training, collaboration, user-centred design, documentation, access, and sustainability. Based on these factors, and inspired by Berners-Lee's (2010) five-star model for Linked Open Data, I propose the following *Five-Star Model for Linked Humanities Data Usability*, aimed at project leaders:

★ **Transparency:** provide clear documentation about the tool/resource, its data structures, and functionality;

★★ **Extensibility:** build upon existing systems or facilitate future extension; encourage integration of new data;

★★★ **Intuitiveness:** develop clear user journeys to facilitate completion of intended tasks;

★★★★ **Reliability:** ensure consistent functionality, while minimising downtime;

★★★★★ **Sustainability:** support continued functionality for (at least) a fixed period.

The above components appear in the order in which they might be considered when planning production of a Linked Humanities Data tool or resource; however, during development they are likely to be addressed simultaneously. Throughout the development process, it is crucial for people with requisite skills and knowledge (or sufficient interest to acquire them through training) to work collaboratively. Eventually, these collaborations can form communities of practice, e.g., those that support *Pelagios* and *Papyri.info*. Such communities can assist in maintaining tools or resources in the long term, facilitating development, managing new contributions, and sharing knowledge.

Conclusions

Linked Humanities Data usability is affected by decisions made during the planning stages of tool or resource development; key factors that facilitate this usability, as demonstrated through my *Five-Star Model*, could equally apply to other technological approaches. However, because Linked Data production is relatively complex and because the potential for reuse is so great, particular care must be taken in ensuring its usability by the research community. This framework of key factors to consider early in the development process should help encourage future Linked Humanities Data producers to prioritise usability and identify areas requiring support or training. As a result, a wider audience will be better able to benefit from the advantages of Linked Data technologies.

Bibliography

Barbera, M. (2013). Linked (open) data at web scale: research, social and engineering challenges in the digital humanities. *JLIS.it*, 4(1): 91–101.

Berners-Lee, T. (2010). *Linked Data - Design Issues*. Available at: https://www.w3.org/DesignIssues/LinkedData.html.

Geser, G. (2016). *ARIADNE WP15 Study: Towards a Web of Archaeological Linked Open Data*. ARIADNE. Available at: http://legacy.ariadne-infrastructure.eu/wp-content/uploads/2019/01/ARIADNE_archaeological_LOD_study_10-2016-1.pdf (Accessed: 26 March 2020).

Granados-García, P.L. (2020). *Cultural Contact in Early Roman Spain through Linked Open Data*. PhD. The Open University. Available at: http://oro.open.ac.uk/73887/ (Accessed: 4 March 2021).

Isaksen, L. (2011). *Archaeology and the semantic web*. PhD. University of Southampton. Available at: http://eprints.soton.ac.uk/206421 (Accessed: 24 October 2016).

Linked Art Contributors (no date). *LOUD: Linked Open Usable Data*, *Linked Art*. Available at: https://linked.art/loud/ (Accessed: 21 July 2021).

Ross, S. *et al.* (2015). Building the Bazaar: Enhancing Archaeological Field Recording Through an Open Source Approach. In Wilson, A.T. and Edwards, B. (eds), *Open Source Archaeology: Ethics and Practice*. De Gruyter, pp. 111–29. doi:10.1515/9783110440171-009.

Simon, R. *et al.* (2015). Linking Early Geospatial Documents, One Place at a Time: Annotation of Geographic Documents with Recogito. *e-Perimetron*, 10(2): 49–59.

Wilkinson, M.D. *et al.* (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1). doi:10.1038/sdata.2016.18.

Notes

1. Although usage of digital tools and resources has likely since increased due to the COVID-19 pandemic, the majority of those discussed by participants have not subsequently undergone significant developments. My findings and recommendations should therefore still be applicable.

Exploring a cultural "filter bubble" in artwork databases of two large museums of fine arts

Minster, Sara
bttechsol@gmail.com
Bar Ilan University, Israel

Kizhner, Inna
inna.kizhner@gmail.com
Siberian Federal University, Russian Federation

Zhitomirsky-Geffet, Maayan
maayan.zhitomirsky-geffet@biu.ac.il
Bar Ilan University, Israel

1. Introduction

Digitization of cultural information by GLAM (Galleries, Libraries, Archives and Museums) institutions opens new opportunities for dissemination of cultural data to heterogenous audiences. However, selective information dissemination related to biases in physical collections, policies or difficulties of digitization and aggregation (Mak 2014; Bode 2020; Zhitomirsky-Geffet and Hajibayova 2020; Kizhner et al 2021; Ortolja-Baird and Nyhan 2021) may lead to the creation of a global cultural "filter bubble" where much of the world cultural heritage remains concealed from the public view, thus missing the opportunity to correct the historical injustices and bias in cultural knowledge representation. A filter bubble is a situation when users are only exposed to homogenous information that conforms with their views and prejudices (Pariser 2011). We can say that a cultural "filter bubble" created by culturally and geographically homogenous artwork collections published online encircles users within the information from the dominant culture, thus narrowing their intellectual and cultural horizons (Zhitomirsky-Geffet 2019).

In this study, we aim to investigate and quantitatively measure the cultural "filter bubble" by comparing data publishing and dissemination practices of two famous national museums of fine arts, the Metropolitan Museum of Art, New York, and the Rijksmuseum, Amsterdam. Both museums present artworks to 'contemporary national and international audiences [1]' and appear in a study of the most influential museums published in 2017 (Van Riel and Heijndijk 2017). The study's contribution to bias or "filter bubble" detection in large cultural heritage databases is crucial to many research projects in digital humanities that are based on the analysis of such databases and large amounts of cultural heritage data.

2. Research methodology

The study examined three dissemination channels used by the two museums: 1) data extracted from searchable online collections [2], 2) data extracted from the museum datasets available for the general use via open APIs of both museums as CSV files for the Metropolitan Museum [3] and an XML file for the Rijksmuseum [4], and 3) items with a Collection Property (P195) retrieved from Wikidata using SPARQL query language. As a result, we obtained six databases (three databases for each of the two museums under study, as shown in Table 1). To measure the geographical and cultural diversity of the databases as conveyed by the five criteria of the ethical evaluation framework (Zhitomirsky-Geffet and Hajibayova 2020) adopted in this study, we ran queries in the six databases that computed the distribution of artworks according to the following variables: artworks' continent, country of origin, artist's nationality, and culture.

3. Results

Our results demonstrate that all six databases from both museums focus on the Western culture (Western and Central Europe and North America), although the Metropolitan Museum of Art shows higher cultural diversity scores compared to Rijksmuseum as it also covers ancient cultures, such as Egypt or Mesopotamia (in line with an art historical canon), in its API dataset (Figure 1). Surprisingly, Metropolitan Museum's Wikidata collection focuses on the United States with 90% of all artworks, while almost all the items in the Rijksmuseum Wikidata collection were created in Europe.

We found that Asian cultures are weakly pronounced in all six datasets (16-18% in Metropolitan's API dataset and online search system, less than 5% for all Rijksmuseum's databases, and almost none in both museums' Wikidata collections) compared to Western and Central European cultures (with over 50% for Rijksmuseum's databases and Metropolitan's online database) and North American cultures (with over 20% in all Metropolitan's databases). Japanese culture, with the highest amount of records from Asia, accounts for 4.7% of the records published online by Metropolitan and 2.3% of records accessed via the Rijksmuseum API, with other Asian countries presented to a much lesser extent. In addition, we found that the representation of the native peoples' artworks (mostly from Asia) that the Netherlands conquered in early modern history, such as Dutch East Indies, Java, Batavia or Indonesia, ranges between 0.6% to 2.8% in the three Rijksmuseum databases. Similarly, the artworks of Native Americans constituted only less than 1% of the three Metropolitan's databases. Although our findings may reflect the rates of physical collections of these museums, they do not comply with the ethical evaluation criteria, nor do they

reflect the museums' mission statements, e.g. "Metropolitan collects, studies, conserves, and presents significant works of art across all times and cultures [5]", "Rijskmuseum offers a representative overview of Dutch art and history … as well as major aspects of European and Asian art [6]".

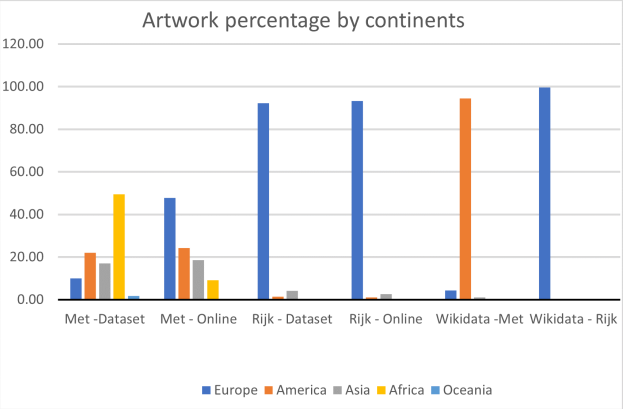


Figure 1
Percentage of items for each continent in all the databases.

Interestingly, the dominance of the Western culture in Wikidata is also reflected by geographic distribution of contributing institutions. Thus, out of the 5,040 (as of June 2021) museums and galleries from around the globe that published their artwork data in Wikidata, 42.25% are located in Western and Central Europe and 17.92% are in North America, while leaving less than 40% to the rest of the world. Only 19.07% of Wikidata museums are located in Asia and 11.93% are in Eastern Europe.

Database name	**Online system**	API Dataset	Wikidata
Metropolitan	516,458	474,383	22,586
Rijksmuseum	716,308	656,665	6,009

Table 1 - Amounts of museum items in each database.

4. Conclusion

We provide evidence to the existence of a cultural "filter bubble" (namely, bias towards Western cultures) in the online databases of the two influential museums, accessed via website search, APIs or queried from Wikidata. All databases appear to have significant constraints and biases in terms of presenting the diversity of cultures, especially in the data submitted to Wikidata, a channel that is important for the dissemination of cultural content among those users who rarely visit institutional websites and museums (Navarrete and Villaespesa 2020). Further investigation into the reasons for the obtained results is needed that could include underrepresentation of various cultures in physical museum collections, curatorial decisions on selective digitization and publishing policies influenced by art history canons or previous institutional policies.

Bibliography

Bode, K. (2020). Why you can't model away bias, *Modern Language Quarterly*, 81: 1.

Inna Kizhner, Melissa Terras, Maxim Rumyantsev, Valentina Khokhlova, Elisaveta Demeshkova, Ivan Rudov, Julia Afanasieva, Digital cultural colonialism: measuring bias in aggregated digitized content held in Google Arts and Culture, *Digital Scholarship in the Humanities*, Volume 36, Issue 3, September 2021, Pages 607–640.

Mak, B. (2014). Archaeology of a digitization. *Journal of the Association for Information Science and Technology*, 65(8): 1515–26.

Navarrete, T., & Villaespesa, E. (2020). Digital Heritage Consumption: The case of the Metropolitan Museum of Art. Magazén, 1(2).

Ortolja-Baird, A. & Julianne Nyhan, (2021). Encoding the haunting of an object catalogue: on the potential of digital technologies to perpetuate or subvert the silence and bias of the early-modern archive, *Digital Scholarship in the Humanities*, 2021; fqab065.

Pariser, E. (2011). The filter bubble: What the Internet is hiding from you. Penguin UK.

Van Riel C., Heijndijk P. (2017). Why people love art museums: a reputation study about the 18 most famous museums among visitors in 10 countries. Rotterdam School of Management, Erasmus University.

Zhitomirsky-Geffet, M. and Hajibayova, L. (2020), "A new framework for ethical creation and evaluation of multi-perspective knowledge organization systems", *Journal of Documentation*, Vol. 76 No. 6, pp. 1459-1471. https://doi.org/10.1108/JD-04-2020-0053.

Zhitomirsky-Geffet M. (2019). Towards a Diversified Knowledge Organization System – An Open Network of Inter-Linked Subsystems with Multiple Validity Scopes. *Journal of Documentation, 75(5):* 1124-1138 .

Notes

1. https://www.rijksmuseum.nl/en/about-us/what-we-do/vision-and-mission
2. https://www.metmuseum.org/art/collection, https://www.rijksmuseum.nl/en/search
3. https://github.com/metmuseum/openaccess/commit/420254f751b4c9d55a3ea7d1dab2d5d1e74e5255#diff-9f9583202c5d326e17789ac08f06b9ec913a7c546a4ab5f68dc32fa9f3732d66
4. https://old.datahub.io/dataset/rijksmuseum-api

5. https://www.metmuseum.org/about-the-met
6. https://www.rijksmuseum.nl/en/about-us/what-we-do/vision-and-mission

EthicsBot: Provoking Ethical Reflection on AI

Mousavi, Emad
emousavi@ualberta.ca
University of Alberta, Canada

Verdini, Paolo
verdini@ualberta.ca
University of Alberta, Canada

Wang, Jingwei
jingwei2@ualberta.ca
University of Alberta, Canada

Barnard, Sara
sbarnard@ualberta.ca
University of Alberta, Canada

Rockwell, Geoffrey
grockwel@ualberta.ca
University of Alberta, Canada

Introduction

A conversational agent (chatbot) is an artificial intelligent software which can simulate a conversation with a user in natural language. With the rapid advances of technology especially in the field of artificial intelligence, chatbots are becoming ever more present in our everyday lives often with a focus on being more and more human-like and assisting their users in everyday activities. In this paper, we will present the EthicsBot project that developed a set of chatbots meant to provoke reflection on the ethics of AI by generating provocative ethical statements in response to input. Our paper will consist of a combination of reflections on how a tool might do ethics and demonstrations of the experimental chatbots developed. In particular we will:

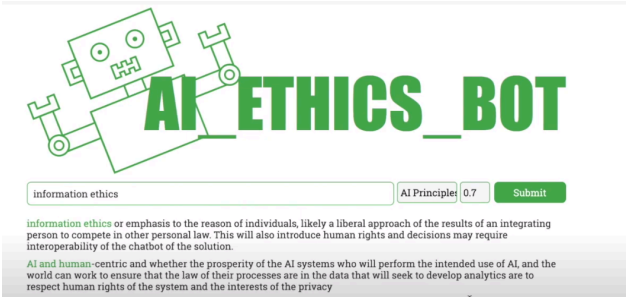
- We will start by reflecting on what it means to build tools to assist in ethics.
- Then we will demonstrate two versions of the EthicsBot that we built.
- Finally we will discuss how we are assessing the output of the EthicsBot.

Reflections on AI ethics

As the title of Mittelstadt 2019 paper argues, "Principles alone cannot guarantee ethical AI." He and others have documented the explosion of AI Ethics principles published by different types of organizations over the last years. These range from those published by Google after the Google Duplex controversy (Griffin, 2018) to the Rome Call on AI Ethics (Rome Call For AI Ethics, 2021).

The genesis of the EthicsBot was an experiment with training a machine with all these sets of AI ethics principles to see whether it might generate provocative new statements following the playful example of Janelle Shane of AI Weirdness (Shane, 2018). Our thinking was that an interface that generated statements based on a seed issue would help the user reflect on that issue much as various thinking tools (De Bono 2015, Ramsay 2011) purport to do. The underlying assumption is that an important part of getting beyond principle to doing ethics is the thinking through of different positions or statements about an ethical issue. Even if the responses were random, bordering on nonsensical, at times the human tendency to make sense (Fyfe, et al, 2008) would benefit the person seeking to think something through. In fact, an AI that was not perfect would itself be an advantage as it would illustrate for the interlocutor the puzzling inconsistency of machine learning. This raised the question of how would we know if the answers generated were, in fact, provocative of reflection, grammatical or not?

Speculative interfaces



In order to train an ethics bot we used a list of AI ethics principles in Hagendorff (Hagendorff, 2020) and Mittelstadt (Mittelstadt, 2019) to gather a corpus. Finding this dataset rather dry we added a second collection of open source "giant brain" type science fiction novels from the Internet

Archive. In the presentation we will show two different interfaces we developed. The first was meant as a public web site (a demo of which could be reached here: https://www.youtube.com/watch?v=vDIMDAInPy8) where users could enter a seed, generate statements, and then select those to keep in a public transcript. The second is a Google Colab notebook that exposes the code so that others can adapt the code. In the end the best results came from the chatbot that used the Generative Pre-trained Transformer 2 - GPT-2 - (GPT-2, 2019) model and our texts on ethics in artificial intelligence. We will review how the code works and go over some of the typical responses it generates depending on the input.

What is in a provocation?

We will then return to a variant of our earlier question: can the EthicsBot provoke ethical thought? To explore this question, we are now curating a set of responses and have asked a panel of experts in the field of ethics in artificial intelligence to assess the responses generated by the EthicsBot based using a rubric similar to what would be used for assessing responses of undergraduate philosophy students in an ethics course. The idea is to compare the responses of the ethics bot to short answer exam questions of the sort "Discuss whether AIs should be regulated?" In principle, if the EthicsBot can approximate an undergraduate quick answer to a seed issue, then it should be capable of responses that could further a user's ethical thinking. What, however, are we assuming about the doing of ethics?

We will conclude the paper by sharing a compilation of the expert panel's assessments of the EthicsBot's responses and their reflections on the usefulness of the chatbot. A sample of how we are gathering the responses from our panel of experts can be reached here: https://forms.gle/TRVx43awwYAwaVKBA

Bibliography

Bono, D. E. (2015). *Lateral Thinking: Creativity Step by Step* (Reissue ed.). Harper Colophon.

Fyfe, S., Williams, C., Mason, O., & Pickup, G. (2008). Apophenia, theory of mind and schizotypy: Perceiving meaning and intentionality in randomness. *Cortex*, *44*(10), 1316-1325. doi: 10.1016/j.cortex.2007.07.009

Griffin, A. (2018). *Google Duplex: Why people are so terrified by new human-sounding robot assistant*. The Independent. https://www.independent.co.uk/life-style/gadgets-and-tech/news/google-duplex-why-explained-controversy-objections-ai-artificial-intelligence-robot-a8347566.html

GPT-2: 1.5B Release. (2019). Retrieved 5 December 2021, from https://openai.com/blog/gpt-2-1-5b-release/

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Mittelstadt, B. (2019). "Principles alone cannot guarantee ethical AI." *Nature Machine Intelligence.* 1: 501-507.

Ramsay, S. (2011). *Reading Machines: Toward an Algorithmic Criticism (Topics in the Digital Humanities)* (1st ed.). University of Illinois Press.

Rome Call For AI Ethics. (2021). Retrieved 18 November 2021, from https://www.romecall.org/

Shane, J. (2018). Opinion | Ruth Bader Hat Guy? Let Our Algorithm Choose Your Halloween Costume (Published 2018). Retrieved 11 December 2021, from https://www.nytimes.com/interactive/2018/10/26/opinion/halloween-spooky-costumes-machine-learning-generator.html

Digital Resistance to Asian-American Hate during COVID-19: Study of Photography and Art on Instagram

Nanditha, Narayanamoorthy
nanditha@yorku.ca
York University

Introduction

In this research, I study the digital resistance to Asian American hate, isolation, alienation, and 'othering' visibilized during the COVID-19 pandemic in 2020-21 in the Global North. Specifically, I draw attention to the role of personal and artistic representations of Asian female bodies that perform both a resistance to hate, in the context of the pandemic, and an affirmation of ethnic and racial heritage and belonging of the self in North America. Through the engagement with #stopasianhate and #haterisavirus hashtags on Instagram, I uncover the rejection of historic and contemporary racial and gendered violence, harassment, xenophobia, and othering that emerges through visual activism and personal and artistic performativity online.

Anti-Asian Hate during the Pandemic

The global spread of COVID-19 has exacerbated hate crimes, and discriminatory acts against Asian Americans who have been "burdened with mounting anxieties and heightened racial tensions, microaggressions, verbal

attacks, physical violence and harassment" (Gover et al 648). According to the Stop AAPI Hate campaign, more than 1700 anti-Asian hate crimes have been documented since the beginning of the pandemic (Jeung). Asian-Americans have been perceived as dangerous (Gover et al 648) owing to the racialized origins of the coronavirus in Wuhan, China, and the "COVID-19 crisis has been misappropriated to reinforce racial discrimination and anti-immigrant rhetoric" (Devakumar et al.). Several incidents from "boycotting of Asian restaurants, bullying of Asian American children at school, to verbal and physical assaults of Asian Americans in public spaces" have been documented during the pandemic (748 - 749). Women and the elderly are most affected by this pervasive violence and the interlinkages between microaggressions and macro structures of power produced by colonialism and white supremacy. The recent mass shooting of six Asian women in Atlanta by Robert Aaron Long, a 21-year-old white old man brought to the fore the risk of physical assault and battery that female Asian bodies carry. In this context, therefore, women have found a safe space on digital platforms to engage in discussions of racism, hate crimes, gendered violence, and participate in a process of sharing in order to subsequently de-'otherize' their own bodies and legitimize their belonging and cultural heritage through digital photography and art.

Method

I conducted this research using visual qualitative analysis of Instagram posts collected on the anti-Asian violence movement. I engaged specifically with #stopasianhate and #hateisavirus hashtags on Instagram using a critical feminist framework and decolonial visual praxis. Out of 535,924, and 51,671 posts yielded by #stopasianhate, and #hateisavirus respectively on Instagram, I manually selected 80 images that depict the self-representation of Asian women, as well as artistic designs for the movement made for/by Asian women in the Global North. To collect Instagram content on the #stopasianhate movement, I manually screenshotted Instagram posts to include the image. All original posts are in English and were collected in the timeline of seven months between April 1st and October 8th, 2021. Here, to understand the relationship between Asian American body politics and digital resistance, I examine various images that showcase the multiplicity of voices in Asian American bodies. With respect to artistic representations of Asian-ness during the pandemic, I investigate the work of multidisciplinary artist Amanda Phingbodhipakkiya (@alonglastname) with her permission.

Argument

Out of the 80 images collected, 34 are exclusively characterized as photographs, self-portraits, or photoshoots that depict the defiance of female Asian American bodies, and the desire to reclaim public and private spaces through their 'Asian-ness.' I contend that digital resistance in photographic and personal representation emerges on Instagram in various ways including as resistance against Asian American hate, opposition to stereotypes of 'Chinese-ness' during the pandemic, navigation of personal identity, and through intergenerational memories of belonging and unbelonging. The process of appropriation, infiltration, and self-insertion of Asian American women on Instagram "combats racial oppression," (Baer 19) produces dissent and disrupts dominant infrastructures. Their bodies emerge as the site of resistance and are constantly being reshaped and repurposed to define their identities and assert their belonging in the North. The visuality of their bodies online creates a counter-hegemonic space against normative hierarchies and power structures where they can produce an agency over and alternate interpretations of their own stories and perform resistance. Their corporeal presence on Instagram and engagement with the #stopasianhate hashtags on digital spaces underpins a sort of precarity, vulnerability, and comfort for these women (Sliwinska 9). The material reconfiguration of public and digital spaces, in fact, allows them to recenter their personal stories through the practice of vulnerable sharing. In this case, photography becomes a powerful medium of creating and disseminating narratives of violent and xenophobic othering.

With respect to artistic representations, Amanda's art in both the physical spaces of New York City as well as on Instagram become a "ritualistic space for self-care and catharsis"; a space that implies ancestral protection; a place to heal as a community from personal grief and collective trauma derived from historical and violent oppression. She brings her artistic creations to life in order to connect with others who identify as Asian Americans while providing a space for community building. Her art performs reclamation of public space and begins the arduous process of re-centering the Asian American narratives. Therefore, visual activism through #stopasianhate, and #hateisavirus in the context of the pandemic, as well as the use of Asian American women's personal and artistic imaginaries on Instagram is a veritable act of decolonization that emerges in the digital resistance of hate. A decolonial praxis in the form of anti-racist hashtags here illuminates the previously invisible narratives of the Asian American community and amplifies the voice of the racialized and gendered subject.

Bibliography

Baer. H. "Redoing feminism: digital activism, body politics, and neoliberalism." *Feminist Media Studies*, 16:1, 17-34, 2016, DOI: 10.1080/14680777.2015.1093070

Devakumar, Delan, et al. "COVID-19: The Great Unequaliser." *Journal of the Royal Society of*

Medicine, vol. 113, no. 6, June 2020, pp. 234–235, doi:10.1177/0141076820925434.

Gover, Angela R., et al. "Anti-Asian Hate Crime During the COVID-19 Pandemic: Exploring the Reproduction of Inequality." *American Journal of Criminal Justice*, vol. 45, no. 4, 2020, pp. 647–67, doi: 10.1007/s12103-020-09545-1.

Jeung, R. "Incidents of coronavirus discrimination march 26-April 1, 2020: A report for A3PCON and CAA." *Asian Pacific Policy and Planning Council*. 2020. Retrieved from http://www.asianpacificpolicyandplanningcouncil.org/wp-content/uploads/Stop_AAPI_Hate_Weekly_Report_4_3_20.pdf.

Sliwinska, Basia. *Feminist Visual Activism and the Body*. Routledge, 2021.

#METOO IN INDIA: MISOGYNY AND THE EMERGENCE OF THE MEN'S RIGHTS MOVEMENT ON TWITTER

Nanditha, Narayanamoorthy
nanditha@yorku.ca
York University

Background

The #MeToo movement in India, as a manifestation of the global #MeToo has invested in the empowerment of Indian women since its inception in October 2017. The digital movement sent reverberations across the country in unprecedented ways and took off on a massive scale in October 2018 (Mathur 2018). The movement constructed around the central hashtag, #MeTooIndia, has successfully created a much-needed discourse on sexual abuse, and harassment at the intersection of sex, power and politics on Twitter. The digital public sphere has facilitated feminist experiences of 'coming-out' with personal stories and experiences for the urban Indian woman.

Research Problem

Cyberspaces in the South are complex and have a different set of constraints from the North, particularly for women and other marginalized collectives. In the discussion of the cyber-South, Gajjala and Oh pose the most important questions: "Will women of the South be allowed or able to use technologies under the conditions that are contextually empowering to them? Within which internet-based contexts can the women of the South truly be heard (2012, 8) or seek to be independent? "Hierarchies of power are deeply embedded in Internet culture as a form of invisible control" (15) of digital feminist spaces of the Global South. Furthermore, the issue of who speaks for whom (Spivak 1988) becomes even more important owing to lack of access, resources, voluntary participation, and representation of a significantly large population.

This is apparent in the discourse surrounding the #MeTooIndia movement where heteropatriarchal and masculinist Hindutva activists organize campaigns of harassment and misogyny against women through a call for men's rights activism that infiltrates the Indian digital feminist movement. Digital platforms in India are becoming increasingly violent, misogynistic, and sexist spaces that enable the congregation of far-right communities that further the everyday offline trauma that women face in the Global South. The presence of hate speech, violence, and masculine toxicity directed towards women acts as a deterrent in participation of women online and in their coming-out in public without anonymity. In addition, this creates unsafe spaces for women and other gendered minorities to recount their experiences with sexual abuse. As feminist activism is becoming increasingly visible on social media platforms; as feminist communities expand and are re-imagined through the use of new media (Mendes et al. 2019, 1), it must be noted that digital culture can be an incredibly complex and toxic space for gendered bodies that are constantly vilified, and objectified. In the context of #MeTooIndia, participants recede into toxic behaviour and vilify the movement for its fake cases, and accusations, or trivialize it for its lack of legal process.

According to authors Vickery and Everback, "mediated misogyny" (2018) oftentimes deliberately infiltrates feminist movements to incite violence, hate, and toxicity within the movement and directed towards members. Serisier's work on rape culture also engages with the emergent culture of speaking out effectively about sexual abuse and rape using online platforms, while underscoring that the act of speaking out can be attributed to certain privileged voices (2007). Under what conditions are we as a collective authorized to speak out, who is receiving, listening, and verifying are important questions that any digital feminist movement needs to consider (2007).

Research Question

This paper explores the emergence of a misogynistic Men's Rights Movement surrounding the #MeToo in India on the social media platform Twitter that leads to gendered exclusion and silencing of feminist collectives in the name of free speech. Using discourse analysis, inductive coding, and close-reading of the corpus of tweets, the research asks the following questions – Does the discourse surrounding #MeTooIndia demonstrate mediated hate speech, violence against women, misogyny, and an emerging men's rights activism? If so, what is the language of violence against women? How do the far-right heteropatriarchal societies on social media employ assertions of independence in their call for men's rights activism? How are feminist digital

movements affected by the hijacking and appropriation by men's rights activists.

Relevance

Although feminist social activism, particularly in the Global South, benefits from Internet culture, and its ability to enable the construction of safe feminist subaltern counterpublics (Fraser 1991) for self-expression, congregation and interaction, Indian women, particularly from marginalized communities, who employ the hashtag #MeTooIndia to participate in the discourse surrounding sexual abuse in India are left feeling unsafe, alienated, isolated, and silenced on social media platforms. The study of the emergence of men's rights activism and misogyny in this context is significant because the discourse around #MeTooIndia, 'hijacked' and targeted by men's rights activists, oftentimes drowns the feminist voice within the movement, leads to the effective erasure of feminist struggles, and creates barriers in participation on digital spaces. In keeping with the theme of the conference, this study also speaks to the congregation of far-right heteropatriarchal societies on social media that employ assertions of independence, and free speech to not merely engage in harassment campaigns against supporters of #MeToo, but also in their call for men's rights activism in India.

Method

10,000 unique tweets were collected and filtered through Twitter Web API between October 1 st, 2018 and October 3rd, 2019 surrounding the #MeToo movement in India. All tweets were collected in the English language to maintain methodological consistency, and retweets were excluded from the sample dataset. Hashtags employed as filters include #metooindia, #indiametoo, #LoSHA, #womanhood +#metooindia, #womanhood+#indiametoo, #sisterhood +#metooindia, and #sisterhood+#indiametoo. Hashtags specifically used in the Indian context have been selected. All tweets are manually annotated and labelled for the specific criterion (below) that is marked on a binary scale of 0/1 where 'Yes' denotes 1, and 'No' denotes 0 -

1. Does this tweet indicate misogyny, violence against women, hate speech or advocation for men's rights activism?

In addition to the inductive coding method, this research employs the critical discourse analysis framework as well as close-reading of tweets to further understand how feminist spaces are dominated by a call for men's rights activism, how heteropatriarchal societies employ assertions of independence on Twitter, and how digital feminist movements are impacted by the hijacking of their spaces in the Indian context. As this research is ongoing, the specific number of tweets that indicate misogyny and the emergence of men's rights activism are pending.

Bibliography

Gajjala, Radhika, and Yeon Ju Oh. 2012. *Cyberfeminism* 2.0. New York: Peter Lang. Print.

Fraser, Nancy. 1990. "Rethinking the Public Sphere: A Contribution to the Critique of

Actually Existing Democracy." Print.

Mathur, Swati. 2018. "India most vocal about #MeToo in October: Global data

analytics co." The Times of India. November 1 https://timesofindia.indiatimes.com/india/india-most-vocal-about-metoo-in-october-global-data-analytics-co/articleshow/66452967.cms.

Mendes, Kaitlynn, Jessica Ringrose, and Jessalynn Keller. 2019. *Digital Feminist*

Activism: Girls and Women Fight Back Against Rape Culture. Kettering: Oxford University Press.

Serisier, Tanya. 2007. "Speaking Out against Rape: Feminist (Her)stories and Anti-rape Politics." *Lilith: A Feminist History Journal* (16): 84–95.

Spivak, G. 1988. "Can the Subaltern Speak?" In Marxism and the Interpretation Culture, edited by C. Nelsson and L. Grossberg's (66–111). Urbana: University of Illinois Press.

Vickery, Jacqueline, and Tracy Everback. 2018. *Mediating Misogyny: Gender, Technology, and Harassment*. Cham, Switzerland: Palgrave Macmillan.

Are Digital Humanities platforms sufficiently facilitating diversity in research? *A study of Transkribus free processing requests.*

Nockels, Joseph Hiliary
j.h.nockels@sms.ed.ac.uk
University of Edinburgh

Terras, Melissa
m.terras@ed.ac.uk
University of Edinburgh

Gooding, Paul
paul.gooding@glasgow.ac.uk

University of Glasgow

Muehlberger, Guenter

guenter.muehlberger@uibk.ac.at
University of Innsbruck

Stauder, Andy

a.stauder@read.coop
Recognition and Enrichment of Archival Documents

Introduction

This paper examines whether free processing initiatives are truly supporting and facilitating research among early career researchers (ECRs), students and those with a lack of funding, as current observations, concerning who is applying for such schemes, suggest that further clarification is often needed. Financial support is necessary to increase the diversity of work seen in the digital humanities (DH), where access to platforms is still unequal. The findings presented in this paper will address several related questions. What demographics are making use of free processing from software developers? What work will this enable? Are schemes being utilised by the intended groups? How can more equitable access be reached?

Research Context

This paper will focus on one software: Transkribus, a handwritten text recognition (HTR) platform which has broadened user access to historical collections through automatic image-to-text recognition, resulting in plain text files which can be presented in a variety of formats by content-holding institutions for instance (Muehlberger et. al, 2019). Current work using Transkribus showcases the breadth of digital humanities research, with models being trained on the manuscripts of Jeremy Bentham (National Library of Scotland, 2021), materials from Ethiopia and Eritrea (Universitat Hamburg, 2021) as well as 18 th -19 th century Bengali print, documents written in Malayalam, and 19th century Devanagari scripts (READ-COOP, 2021).

Since 2020, Transkribus has been developed by the READ-COOP, a cooperative of currently 82 institutions and 20 individual researchers, becoming a paid-for service in October of 2020. With this recent change, a gap has emerged in understanding who is using the tool and what research is being conducted. Alongside this, no systematic review of free processing user requests has been carried out. With the software no longer supplying no-cost text recognition, READ began a free processing scheme for students and those carrying out workshops under the "Transkribus Scholarship Programme" (Transkribus, 2021). Those applying must fill out an online form, indicating the amount of credits needed, their home institution and details of their work. If accepted, credits are then added to the system and a notification is sent. This move to a funded model, with limited free access, offers a glimpse at how software companies are balancing sustainability with ensuring as equitable access as possible to their products.

There is a historical issue of access inequality for DH tools and infrastructures. This goes beyond cost, raising issues of language, hardware requirements and barriers to entry in terms of computational knowledge, raising a bigger problem of limiting access to culture (Spiro, 2011: 1-10). As such, an unhealthy weighting toward Global North insights has occurred within the field (Risam, 2015: 161-175). This requires a social justice minded approach, designing new workflows and tools which resist previous inequalities.

That said, due to the nature of the processing requests, this paper looks at access in terms in funding, aiming to answer whether free processing schemes can be part of a social justice approach toward making DH platforms more equitable. Using a pay-for model can easily create marginalisation for certain user groups. Though hard to define precisely across institutions and nationalities, this study will focus on students, those completing degree awards; ECRs, those who are engaging in post-doctoral research and transitioning toward being independent academics (UKRI, 2020), and those who lack funding for their work. Though students are easily identifiable through these requests, ECRs are harder to ascertain - despite many offering only details of their position when writing about their research. In the case of missing data, world rankings will be used to detail the institutional income of those making requests (QS World Rankings, 2022), while strategies for further engagement are developed.

This paper looks to fill a gap in understanding concerning who is benefiting from this free processing scheme through a systematic review of online requests. In turn, a glimpse at how HTR tools are currently being made accessible will be reached. Whether current schemes are truly facilitating diverse research, or ignoring existing inequalities in the field, will also be cited.

Methodology

Content analysis will be applied to these processing requests, alongside interviewing READ staff. While these online requests vary in detail, they provide data on: the required number of processing credits; the discipline and institution of the user; the user's current academic position; and a short project description. This study will examine over 150 requests collected between November 2020 and March 2022. These requests will be aggregated and anonymised, in accordance with gained ethics approval from the University of Edinburgh. They will then be interrogated using content analysis, a research technique for the "objective, systematic and quantitative description of manifest content of communication" (Harvey, 2020). These requests will be coded, capturing their contents, using a mix of in vivo codes, quoted directly from the data (Saldana,

2012: 10), and process codes to gain a sense of the actions users are completing with Trankribus (Corbin and Strauss, 2015: 283). Through this method, this research will apply a set of procedures to make valid inferences from free-processing requests, presenting replicable and valid results (Krippendorff, 1980: 71) as to whether current schemes are weakening the financial barriers being faced by users of Transkribus.

The information gained from reviewing the free processing requests will sit alongside information from interviews of READ staff, ascertaining what influenced the decision to supply free processing to these groups and what the aims were.

Conclusion

This paper explores the extent to which Transkribus supports early career and marginalised scholars in accessing the platform, using content analysis of free processing requests and interviews with members of the Transkribus staff. As the first study to systematically analysis these requests, it provides important insights into how the transition to a paid-for model has impacted Transkribus's users. It provides insights into the demographics of users requiring free processing, the types of projects which are being supported, and how successful the READ-COOP has been in supporting research. These findings allow us to develop recommendations for improving access to Transkribus, as well as begin to draw parallels to other HTR providers in making these platforms more equitable.

Bibliography

Corbin, J., Strauss, A. 2015. Basics of Qualitative Research. Thousand Oaks, CA: Sage.

Harvey, L. 2020. Content Analysis, Social Research Glossary, Quality Research International. https://www.qualityresearchinternational.com/socialresearch/ . Accessed June 1, 2021.

Krippendorff, K. 1980. Validity in Content Analysis. In E. Mochmann, E. (ed.), *Computerstrategien* für *die kommunikationsanalyse*. Frankfurt, Germany: University of Frankfurt Press, pp. 69-101.

Muehlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., Colutto, S., Dejean, H., Diem, M., Fiel, S., Gatos, B., Greinoecker, A., Gruning, T., Hackl, G., Haukkoyaara, V., Heyer, G., Hirvonen, L., Hodel, T., Jokinen, M., Kahle, P., Kallio, M., Kaplan, F., Kleber, F., Labahn, R., Lang, E.M., Laube, S., Leifert, G., Louloudis, G., McNicholl, R., Meunier, J.L., Michael, J., Muhlbauer, E., Philipp, N., Pratikakis, I., Puigcerver Perez, J., Putz, H., Retsinas, G., Romero, V., Sablatnig, R., Sanchez, J.A., Schofield, P., Sfikas, G., Sieber, C., Stamatopoulos, N., Strauss, T., Terbul, T., Toselli, A.H., Ulreich, B., Villegas, M., Vidal, E., Walcher, J., Weidemann, M., Wurster, H., Zagoris, K. (2019). Transforming scholarship in the archives through handwriting text recognition, Transkribus as a case study. *emerald publishing*, 75(50): 960-970.

National Library of Scotland Data Foundry. 2021. Diaries, letters and poems of Marjory Fleming's diary. https://data.nls.uk/data/digitised-collections/marjory-fleming . Accessed November 16, 2021.

QS World Rankings. 2021. https://www.topuniversities.com/university-rankings/world-university-rankings/2022 . Accessed November 20, 2021.

READ-COOP. 2021. Recognising printed Asian texts with Transkribus. https://readcoop.eu/printed-asian-text/ . Accessed November 16, 2021.

Risam, Roopika. 2015. South Asian Digital Humanities: An Overview. *South Asian Review*, 36(3): 161-175.

Saldana, Johnny. 2012. The Coding Manual for Qualitative Researchers. London: Routledge.

Spiro, Lisa. 2011. Getting Started in Digital Humanities. *Journal of Digital Humanities*, 1(1): 1-10.

Transkribus Scholarship Programme. 2021. https://readcoop.eu/transkribus/scholarship/ . Accessed November 19, 2021.

UK Research and Innovation (UKRI). 2020. Early career researchers: career and skills development. https://www.ukri.org/councils/ahrc/career-and-skills-development/early-career-researchers-career-and-skills', development/. Accessed November 10, 2020.

Universitat Hamburg. 2021. About beta masaheft. https://www.betamasaheft.uni-hamburg.de/about.html . Accessed November 16, 2021.

Small Data projects / Big Data research: contemporary problems and historical solutions

O'Donnell, Daniel
daniel.odonnell@uleth.ca
Humanities Innovation Lab, University of Lethbridge, Canada

Woods, Nathan
nathan.woods@uleth.ca
Humanities Innovation Lab, University of Lethbridge, Canada

Bordalejo, Barbara
barbara.bordalejo@uleth.ca
Humanities Innovation Lab, University of Lethbridge, Canada

Humanist resistance to understanding the material they work with as "data" is well-documented (e.g. Marche, 2012; Fish, 2012). As O'Donnell has argued,

> In other domains, data are generated through experiment, observation, and measurement. Darwin goes to the Galapagos Islands, observes the finches, and fills notebooks with what he sees. His notes (i.e. his "data")... are "the facts, numbers, letters, and symbols that describe an object, idea, condition, situation, or other factors." Given the extent to which they are generated, it has been argued that they might be described better as capta, "taken," than data, "given."
>
> The material of humanities research traditionally is much more datum than captum, finch than note…. [S]uch material... is often unique and its interpretation is usually provisional, depending on broader understandings of purpose, context and form that are themselves open to analysis, argument and modification. In the humanities, we more often end up debating why we think something is a finch than [listing] what we can conclude from observing it (O'Donnell, 2016).

Scale is an important result of this distinction. While experimental and observational approaches to data generation can produce immense datasets, the more dialogic approach taken traditionally by Humanities and Cultural Heritage (HCH) researchers often leads to the development of relatively small, closely analysed datasets or even datapoints — the edition of a single novel or shipping register; a collection of comics; the oeuvre of a single artist or school (Borgman, 2015; Borgman, 2007; Golub and Liu, 2021).

This difference often results in a mismatch between research infrastructures and the needs of many small-data HCH researchers and users. The evolution of Open Science/Scholarship Infrastructure (OSI) offers a case in point. OSI typically assumes a research workflow and understanding of the purpose and nature of data in which data is clearly distinct from analysis and, traditionally, represents the raw material of research rather than a research output in its own right (Flanders, 2009; Jockers and Flanders, 2013). OSI, in this use case, provides a forum for the registration and publication of what was in many cases previously unpublished and considered unpublishable (Gray et al., 2002; Kratz and Strasser, 2015).

This is different from the typical understanding of the role of "data" (whether recognised as such or not by the researchers) in traditional small-data HCH editions and exhibits. In these projects "data," which in this case we are defining as the mediated and curated representation of primary texts and objects intended to be used by others as a proxy for access to the originals, is understood to be a principal research output in its own right. In contrast to the workflow contemplated by contemporary OSI, in which data precedes and is published separately from analysis, in this workflow data is commonly given pride of place: published with the accompanying analysis and intended to be used directly by the end-user. In these cases, "analysis" is, if anything, often treated almost as a form of metadata rather than a distinct set of results derived from an underlying dataset.

This paper explores the question of Research Data Management (RDM) in this context: for small-data projects involving the representation of HCH texts and objects. Our primary focus is the understanding of what we will call "data" (regardless of the views of the researchers themselves about this terminology) inherent in such projects' research and publication workflows (O'Donnell, 2018). We will expand on the distinction drawn above between "capta" and "data" (recognising at the same time the instability of the each term's valency, cf. O'Donnell, 2016 and Drucker, 2011) and focus on how such "representational data" is captured, reproduced, and used in traditional HCH "primary-source" research workflows and use cases such as editions, catalogues, and exhibits.

The paper contrasts this analysis of HCH small-data practices and use cases with the understanding of data and research/RDM workflows implied or described by various specific examples of OSI (e.g. Zenodo, Figshare, Open Science Framework, Humanities Commons). We use this contrast to identify ways in which OSI can be adapted to support such small-data work and make it available for "big data" research (for an example of this adaptation within the sciences themselves, see Ferguson et al., 2014; Seltmann et al., 2013; Biodiversity Literature Repository, 2013; Cui et al., 2010).

This is not the first time HCH researchers have struggled with this problem. An important part of the paper is an exploration of how such adaptations have been made in the past. While the contemporary conversation around big data has focused on features such as quantification, automisation, standard graphical representation of data patterns, and the compilation of large datasets, we build upon recent scholarship in the history of science that reframes the pre-history of this current conversation to consider alternate precedents (Aronova et al., 2010).

Our focus in that case is on HCH projects — such as the Oxford English Dictionary (OED) and the Corpus Inscriptionum Latinarum (Daston, 2017) — that created big data projects built from distributed data networks. As the architects of these projects demonstrated in the 19th century,

it is possible, often with considerable effort, to reuse such work for big-data ends. The OED, after all, is a 'big data' project built on the basis of a collection (i.e. a 'dataset') of 1.8 million quotations collected from thousands of books (i.e. 'small-data' projects; see Oxford English Dictionary; Trench, 1860). These data networks compiled individual quotations and inscriptions as published compendia or editions.

By exploring these big data precedents in the history of the humanities, our paper contributes in critical and practical ways to the contemporary and ongoing discussion on the organisation of data projects in Digital Humanities (Antonijević, 2015; Antonijević Ubois, 2016; Borgman, 2015; Posner, 2013). While recent scholarship has focused on the work of organising big data Digital Humanities projects, we argue this conversation has often conflated issues of RDM with issues of digital practices in knowledge production. In our concluding discussion, we explore how these earlier examples of data management might enrich contemporary discussion, particularly around the development of scholarly tools and research data management infrastructure. In reframing how digital practice and data might be related and combined, we reconsider in practical and historical terms how broader genealogies suggest alternative portraits of how humanities data is compiled, organised, used and shared.

Bibliography

Antonijević, S.(2015). *Amongst Digital Humanists: An Ethnographic Study of Digital Knowledge Production*. First published. Basingstoke New York, NY: Palgrave Macmillan doi:10.1057/9781137484185.

Antonijević Ubois, S.(2016). Developing Research Tools via Voices from the Field Text *Dh+lib*World. https://acrl.ala.org/dh/2016/07/29/developing-research-tools-via-voices-from-the-field/ (accessed 20 April 2022).

Aronova, E., Baker, K. S. and Oreskes, N.(2010). Big Science and Big Data in Biology: From the International Geophysical Year through the International Biological Program to the Long-Term Ecological Research (LTER) Network, 1957–Present. *Historical Studies in the Natural Sciences*, **40**: 183–224 doi:10.1525/hsns.2010.40.2.183.

Biodiversity Literature Repository(2013). *Biodiversity Literature Repository [Project]* https://zenodo.org/record/3475439.

Borgman, C. L.(2007). *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, Mass: MIT Press.

Borgman, C. L.(2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT press https://cloudfront.escholarship.org/dist/prd/content/qt6vt1h4wt/qt6vt1h4wt.pdf.

Cui, H., Jiang, K. (Yang) and Sanyal, P. P.(2010). From text to RDF triple store: An application for biodiversity literature. *Proceedings of the American Society for Information Science and Technology*, **47**(1): 1–2 doi:10.1002/meet.14504701415.

Daston, L.(2017). The Immortal Archive: Nineteenth-Century Science Imagines the Future. *Science in the Archives: Pasts, Presents, Futures*. London and Chicago: UChicago Press.

Drucker, J.(2011). Humanities Approaches to Graphical Display. *Digital Humanities Quarterly*, **005**(1).

Ferguson, A. R., Nielson, J. L., Cragin, M. H., Bandrowski, A. E. and Martone, M. E.(2014). Big data from small data: data-sharing in the 'long tail' of neuroscience. *Nature Neuroscience*, **17**(11). Nature Research: 1442–47 doi:10.1038/nn.3838.

Fish, S.(2012). Mind Your P's and B's: The Digital Humanities and Interpretation *Opinionator* https://bit.ly/3gKM3mC (accessed 30 March 2013).

Flanders, J.(2009). The Productive Unease of 21st-century Digital Scholarship. , **3**(3) http://digitalhumanities.org/dhq/vol/3/3/000055/000055.html (accessed 12 May 2013).

Golub, K. and Liu, Y.-H.(2021). *Information and Knowledge Organisation in Digital Humanities: Global Perspectives*. 1st ed. London: Routledge doi:10.4324/9781003131816. https://www.taylorfrancis.com/books/9781003131816 (accessed 17 January 2022).

Gray, J., Szalay, A. S., Thakar, A. R., Stoughton, C., and others(2002). Online scientific data curation, publication, and archiving. *Virtual Observatories*, vol. 4846. International Society for Optics and Photonics, pp. 103–07.

Jockers, M. and Flanders, J.(2013). A Matter of Scale. *Faculty Publications – Department of English* https://digitalcommons.unl.edu/englishfacpubs/106.

Kratz, J. E. and Strasser, C.(2015). Researcher Perspectives on Publication and Peer Review of Data. *PLOS ONE*, **10**(2). Public Library of Science: e0117619 doi:10.1371/journal.pone.0117619.

Marche, S.(2012). *Literature Is Not Data: Against Digital Humanities - Los Angeles Review of Books*. *Los Angeles Review of Books* https://lareviewofbooks.org/article/literature-is-not-data-against-digital-humanities/.

O'Donnell, D. P.(2016). The bird in hand: Humanities research in the age of Open Data. In Figshare (ed), *The State of Open Data Report*. Digital Science, pp. 38–39 doi:10.6084/m9.figshare.4036398.v1. https://figshare.com/articles/The_State_of_Open_Data_Report/4036398.

O'Donnell, D. P.(2018). Humanities Data and their Research Use Paper presented at the Open Science

Infrastuctures for Big Cultural Data, International Masterclass, Plovdiv, Bulgaria https://zenodo.org/record/2246390 (accessed 14 January 2021).

Oxford English Dictionary *History of the OED*. *Oxford English Dictionary* https://public.oed.com/history/.

Posner, M.(2013). No Half Measures: Overcoming Common Challenges to Doing Digital Humanities in the Library. *Journal of Library Administration*, **53**(1): 43–52 doi:10.1080/01930826.2013.756694.

Seltmann, K. C., Pénzes, Z., Yoder, M. J., Bertone, M. A. and Deans, A. R.(2013). Utilizing Descriptive Statements from the Biodiversity Heritage Library to Expand the Hymenoptera Anatomy Ontology. (Ed.) Moreau, C. S. *PLoS ONE*, **8**(2): e55674 doi:10.1371/journal.pone.0055674.

Trench, R. C.(1860). *On Some Deficiencies in Our English Dictionaries: Being the Substance of Two Papers Read Before the Philological Society, Nov. 5, and Nov. 19, 1857*. *Oxford English Dictionary*. J. W. Parker and son https://public.oed.com/history/archives/on-some-deficiencies/appendix/.

Co-reference networks for dramatic texts: Network analysis of German dramas based on co-referential information

Pagel, Janis
janis.pagel@uni-koeln.de
University of Cologne, Germany

Introduction

Social network analysis (Abraham et al., 2009) plays an important role in the computational analysis and quantification of dramas (Moretti, 2011; Trilcke et al., 2015; Fischer et al., 2017).

Usually, these networks are based on the co-presence of stage characters. In such a network, each node represents a character and edges between nodes indicate if two characters co-occurred on stage, i.e. if they appeared in the same scene.

For other types of literary texts such as novels, there have been several works on what type of information the edges could be based on, such as adjacent quoted speech (Elson et al., 2010), topic frequencies (Celikyilmaz et al., 2010), social events (Agarwal et al., 2012) or similarities of word embeddings (Wohlgenannt et al., 2016). However, for dramatic texts, the majority of research on social networks only uses the co-occurrence of characters as a basis for edges.

In this paper, we propose to not only model when characters interact on stage, but also when a character is mentioned by another character while this character is not present. This follows the approach by Agarwal et al. (2012) and Agarwal et al. (2010) who model different types of social events, including characters mentioning and thinking about one another.

As Agarwal et al. (2012) already point out, many of the social events described in Agarwal et al. (2010) do not normally occur in historical literary works. For dramatic texts, there are further restrictions, e.g. thinking about another character will not occur in dramatic texts, but is a frequent occurrence for the novel *Alice in Wonderland*, used by Agarwal et al. (2012).

We therefore focus on the social event of characters mentioning each other, while the mentioned character is not present on stage. [1] In order to access this information, we make use of a corpus of annotated co-references and speaker tags.

Application of Co-reference Networks

The following networks are constructed on data from 31 plays annotated for co-reference (Pagel & Reiter, 2020). [2] These annotations build upon the TEI-annotations of the DraCor initiative (Fischer et al., 2019), which provides the co-presence information. [3]

Figure 1 shows the application of a co-reference network on Gotthold Ephraim Lessing's

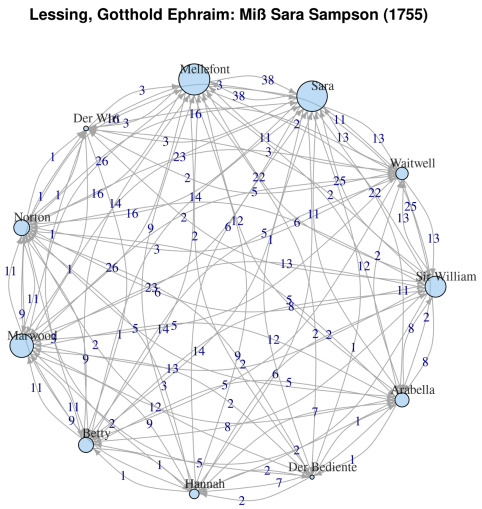


Fig. 1:
Network of Miß Sara Sampson *with co-reference information. Node size is scaled according to weighted degree.*

play *Miß Sara Sampson*. Each node represents a character in the play and each edge represents a mention made by one character about another character. The direction of the

mention is indicated by arrows, while the arrow points towards the character being mentioned. The weights on each edge indicate how often the character in question is mentioned by another character. Nodes are scaled in size according to their weighted degree, which is the sum of all incoming and outgoing weights of a node (Barrat et al., 2004). It can be observed that the three main characters of the play *Mellefont*, *Sara* and *Marwood* receive a higher weighted degree compared to the other characters, showing that the network models character importance to a certain degree, just like a co-presence network would.

Comparison to Co-presence Networks

Comparing the network in Figure 1 to the co-presence network of the same play in Figure 2 reveals commonalities and differences between the two approaches. Again, the three main characters are exponated compared to other side characters, however, the overall weighted degrees are much smaller. It is noticeable that *Sir William*, Sara's father, holds much less importance in the co-presence network, while he has much more weight compared to other characters in the co-reference network in Figure 1. Furthermore, *Marwood* has the same degree value as *Betty*, Sara's maidservant, while in the co-reference network, it becomes clear that Marwood holds a more important role in the social configuration of the play compared to Betty.

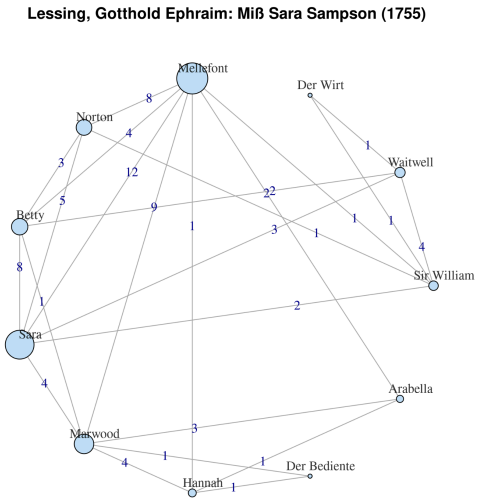


Fig. 2:
Network of Miß Sara Sampson *with co-presence information. Node size according to weighted degree.*

A reader of the play will be able to make several observations: Sir William is often not present on stage, but plays an important role in the plot and in the decision making of Sara and Mellefont. Furthermore, Marwood plays a more important role in the overall plot of the play than Betty does. However, quantitatively, this only really becomes apparent when comparing and juxtaposing the two different networks.

We can also see this complementary nature when abstracting away from looking at the pure networks and comparing the development of weighted degree during the course of the play.

Figure 3 shows the current weighted degree for the three characters *Sara*, *Mellefont* and *Marwood* when constructed from information available for each scene of the play. For example, for Scene 1, a co-presence network with all co-presences of the character in this scene is constructed and the weighted degree is computed based on this network. The same is done for all further scenes and separately for all coreferent mentions of a character. We can clearly see that especially Marwood and Sara are frequently mentioned in scenes from which they are physically absent. [4]

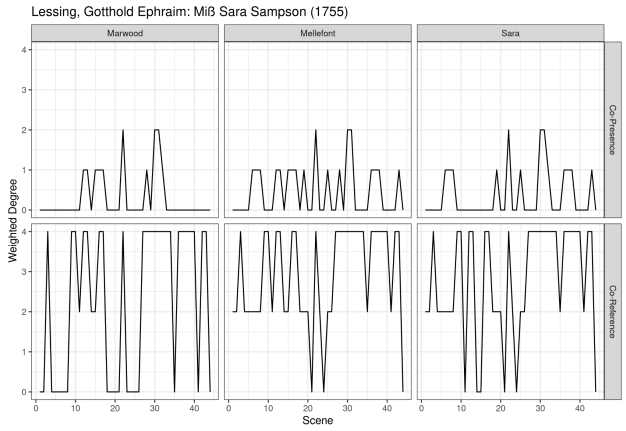


Fig. 3:
Development of weighted degree per scene for co-presence and co-reference in Miß Sara Sampson.

Lastly, when looking at a comparison of all annotated 31 plays, seen in Table 1, the overall averaged values for different centrality measures, namely degree, betweenness, closeness and eigenvector centrality (cf. Newman, 2010, chap. 7), are higher for co-reference networks than for co-presence networks, except for closeness [5]. At the same time, the standard deviations for the measures are comparable. This shows that while capturing the same phenomenon, which is the importance of characters in the social constellations of the plays, individual characters receive higher importance depending on co-presence or co-reference, and as such, the complementary information that both approaches provide can also be seen on a larger scale.

	Centrality Measure	Mean	Standard Deviation
Co-presence	Degree	0.48	0.28
	Betweenness	0.04	0.08
	Closeness	0.44	0.18
	Eigenvector	0.41	0.31
Co-reference	Degree	0.55	0.30
	Betweenness	0.05	0.09
	Closeness	0.43	0.18
	Eigenvector	0.47	0.33

Tab. 1: Mean values for different centrality measures in either the co-presence or co-reference networks and standard deviation.

Conclusion

We have presented the conception and application of social networks for German theatre plays. The approach can easily be applied for dramas of other languages with annotated co-presence and co-reference information. Furthermore, we demonstrated that the currently widely used co-presence networks can be complemented by the use of co-presence networks and it appears to be beneficial to use both types of information when conducting network analysis on larger collections of texts, in order to access different layers of character importance. This was shown both on the level of single plays as well as on the level of a small corpus. One way to quantitatively capture the different perspectives the two types of networks contribute has been shown by using averaged centrality measures. There are also potential shortcomings of the approach, e.g. a comparison of the two networks would not capture characters which are mentioned but never appear on stage. Furthermore, additional approaches of comparing the network types may be developed, for instance comparing the networks more directly by utilising distance measures.

Bibliography

Abraham, A., Hassanien, A.-E., & Snášel, V. (2009). *Computational Social Network Analysis: Trends, Tools and Research Advances*. Springer Science & Business Media.

Agarwal, A., Corvalan, A., Jensen, J., & Rambow, O. (2012). Social Network Analysis of Alice in Wonderland. *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, 88–96. https://aclanthology.org/W12-2513

Agarwal, A., Rambow, O., & Passonneau, R. J. (2010). Annotation Scheme for Social Network Extraction from Text. *Proceedings of the Fourth Linguistic Annotation Workshop*, 20–28. https://aclanthology.org/W10-1803

Barrat, A., Barthélemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences*, *101*(11), 3747–3752. https://doi.org/10.1073/pnas.0400087101

Celikyilmaz, A., Hakkani-Tur, D., He, H., Kondrak, G., & Barbosa, D. (2010). The Actor-Topic Model for Extracting Social Networks in Literary Narrative. *Proceedings of the NIPS 2010 Workshop – Machine Learning for Social Computing*. https://webdocs.cs.ualberta.ca/~denilson/files/publications/nips2010.pdf

Elson, D., Dames, N., & Mckeown, K. (2010). Extracting Social Networks from Literary Fiction. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, 138–147. https://aclanthology.org/P10-1015

Fischer, F., Börner, I., Göbel, M., Hechtl, A., Kittel, C., Milling, C., & Trilcke, P. (2019). Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on

European Drama. *Proceedings of DH2019: "Complexities."* https://zenodo.org/record/4284002

Fischer, F., Göbel, M., Kampkaspar, D., Kittel, C., & Trilcke, P. (2017). Network Dynamics, Plot Analysis. Approaching the Progressive Structuration of Literary Texts. *Book of Abstracts of the DH2017 Conference*.

Moretti, F. (2011). Network Theory, Plot Analysis. *Pamphlets of the Stanford Literary Lab*, *2*, 2–11.

Newman, M. (2010). *Networks: An introduction*. Oxford University Press.

Pagel, J., & Reiter, N. (2020). GerDraCor-Coref: A Coreference Corpus for Dramatic Texts in German. *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 55–64. http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.7.pdf

Trilcke, P., Fischer, F., & Kampkaspar, D. (2015). Digital network analysis of dramatic texts. *DH2015 Conference Abstracts*.

Willand, M., Krautter, B., Pagel, J., & Reiter, N. (2020). Passive Präsenz tragischer Hauptfiguren im Drama. In C. Schöch (Ed.), *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts* (pp. 177–181). https://zenodo.org/record/3666690

Wohlgenannt, G., Chernyak, E., & Ilvovsky, D. (2016). Extracting Social Networks from Literary Text with Word Embedding Tools. *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, 18–25. https://www.aclweb.org/anthology/W16-4004

Notes

1. Called "Cognition Event" in Agarwal et al. (2010).
2. The data used in this paper is available at https://doi.org/10.5281/zenodo.4792311.
3. TEI-XML available from https://github.com/dracor-org/gerdracor
4. This type of observation can also be explored in a comparison of active and passive stage presence, see for example Willand et al. (2020).
5. The fact that closeness is lower for co-reference networks when compared to co-presence networks is to be expected, since networks with a fewer number of edges have a higher chance to receive higher closeness values per node. However, as the difference is rather small, it is difficult to make generalisations about the differences between the two network types with regards to closeness and based on the data.

Knowledge organization of the Hong Kong Martial Arts Living Archive to capture and preserve intangible cultural heritage

Picca, Davide
davide.picca@unil.ch
University of Lausanne - Switzerland

Adamou, Alessandro
adamou@biblhertz.it
Bibliotheca Hertziana - Max Planck Institute for Art History - Italy

Hou, Yumeng
yumeng.hou@epfl.ch
Laboratory for Experimental Museology - EPFL - Switzerland

Egloff, Mattia
mattia.egloff@unil.ch
University of Lausanne - Switzerland

Kenderdine, Sarah
sarah.kenderdine@epfl.ch
Laboratory for Experimental Museology - EPFL - Switzerland

Introduction

Since the 2003 UNESCO convention stated the importance of preserving intangible cultural heritage (ICH), numerous efforts were undertaken to expand traditional cultural heritage to encompass immaterial aspects. However, the boundaries of what should be intended as ICH are not set: the intangible element can be found in such areas as the aesthetics (e.g. performing arts), epistemology (style, semiotics) and transmission (oral history) of culture.

Treating martial arts as cultural heritage potentially incorporates all these categories of immateriality The Hong Kong Martial Arts Living Archive (HKMALA) as a testimony of arts, styles and methods that are passed down onto small communities, is a prominent example of multiple degrees of ICH occurring together. Whilst gathering a mass

of audio-visual and motion capture content to visually preserve endangered Southern Chinese cultures (see Figure it offers an opportunity to turn it into structured knowledge, being originally organized around its past run of exhibitions.

We present an ongoing endeavor to extract and formalize Hakka martial arts knowledge out of HKMALA content into a knowledge graph of linked datasets, thus offering a ground truth for future research questions on the understanding of cultural contact across martial arts communities.

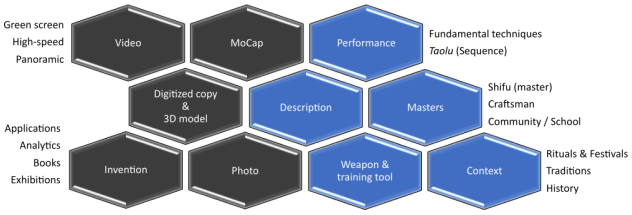


Figure 1:
Overview of HKMALA content types.

A basic ontological framework for martial arts

Barring sports or military contexts, no known ontology presents a unified theory of the martial discipline, to form the basis of such a knowledge organization of archival content. As an initial task, this project built one such framework. This is modelled as an ontology network grounded on foundational ontologies, rather than cultural heritage models, but whose modularity reflects the intent to highlight the aesthetic, epistemic and social traits of martial arts (Figure The identification of which traits to be considered culturally relevant will subsequently be delegated to inference models (rule systems, reasoners) combining domain ontologies with cultural ones like CIDOC-CRM and ArCo See , for details on our ontology engineering work.

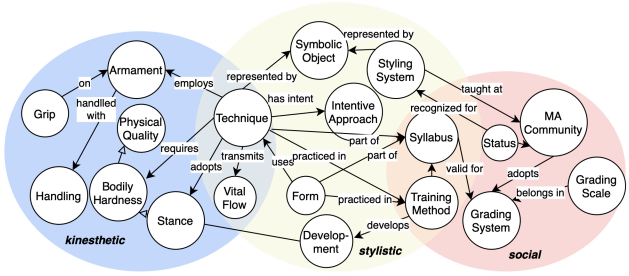


Figure 2:
Documented ontology at https://crossings.github.io/ont/

Instantiation on the HKMALA content

A knowledge organization of HKMALA involves (1) annotating media content of technique demonstrations, MoCap segments and interviews to masters, and (2) indexing these media so that they can be consumed using computational methods like ontology-based data access and semantic query languages like SPARQL. The above ontologies offer a basic framework for a martial arts data domain, yet do not specifically model HKMALA's Hakka Kung Fu domain: therefore, an instantiation effort is required. This is accomplished, as per Figure by:

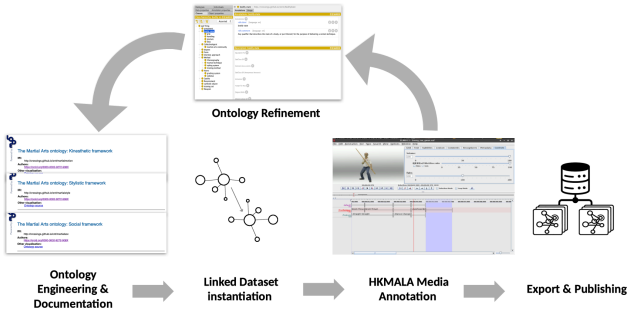


Figure 3:
Pipeline of HKMALA knowledge organization.

1. Constructing the Hakka martial arts dataset from selected sources.
2. Semantically annotating HKMALA media content using terminology lifted from the dataset in question.
3. Iteratively refining our martial arts ontologies to satisfy the versatility required by the media annotation schema.
4. Continuously publishing the knowledge graph resulting from datasets (1,2) as FAIR data

To generate the dataset that instantiates the Martial Arts ontology (1), we performed data extraction and reengineering from the following sources:

a) texts of past HKMALA exhibition panels and captions; b) glossaries in the literature on Southern Chinese martial arts referenced by said texts; c) manifest files containing tabular data to assemble exhibitions out of the archive content; d) transcripts of interviews to masters.

Entity extraction from the texts was performed using the Stanford named entity tagger. Part of these sources had previously been used to bootstrap the ontology itself, therefore the terms not used then that denoted instances were included in the dataset.

Semantic media annotation (2) is performed using the ELAN toolkit, a multi-tiered video annotation software that allows custom terminology and stores annotation in

the open format EAF (which can be exported to JSON-LD and is therefore interoperable with the standards underneath knowledge graphs. Our tiered organization of EAF reflects the three dimensions of cultural heritage that constitute the ontology modules: for example, a video of a Southern Praying Mantis master demonstrating a technique will be annotated along a layer for the aesthetic dimension (e.g. movement, stance and body parts involved), one for the epistemic dimension (technique, style and symbolic reference such as the mantis itself), and one for the social/ transmission dimension (the master's identity).

The scrutiny and annotation of HKMALA media files brought to light a need to refine the original ontological framework (3) to accommodate the expressivity of the knowledge encoded in the medium. For example, a master explaining what qualities are developed by a training exercise on what parts of the body required that an n-ary relation should be materialized as a Development class. Similarly, distinguishing techniques where the energy flow (*qi*) comes from the point of contact (external) or from the attacker's body (internal) hinted that a class FlowTransmissionType should be formalized.

The dataset and annotations (4) are published in formats compliant with the Linked Data standards (i.e. Turtle, N-Triples and JSON-LD), both on GitHub [1] and on Zenodo [2]and following a rolling release / continuous update model. The choice of representation formats and publishing channels was driven by the need to a) ensure data FAIRness and the ability to load them onto any triple store for querying; b) ease of availability for programming libraries to download and use the data in client code, as exemplified in the concluding section.

Conclusions

Aiming at creating a gold standard for building knowledge graphs on martial arts, we offer data and annotations for a subset of the HKMALA content that was originally made publicly available [3], with a view on extending it to the entire archive in the future.

The next step is to align our datasets with the open data cloud: although full third-party coverage is not expected for our domain, Wikidata offers limited authority coverage for some masters, techniques and most importantly geographical and ethnological coordinates.

The project will provide a way to programmatically consume public HKMALA media through an extension of the DHTK Python library, an ontology- based computational model for Humanities data access

Acknowledgement

This work was supported by CROSSINGS - Computational Interoperability for Intangible and Tangible Cultural Heritage, a project in Collaborative Research on Science and Society (CROSS 2021). The authors also acknowledge the *Hong Kong Martial Arts Living Archive*, a research collaboration between the International Guoshu Association, City University of Hong Kong, and the Laboratory of Experimental Museology at EPFL.

Bibliography

Adamou, A., Hou, Y., Picca, D., Egloff, M., Kenderdine, S. (2021). *Ontology- mediated cultural contact detection through motion and style in Southern Chinese martial arts.* In Semantic web and ontology design for cultural heritage (swodch 2021) (Vol. 2949). CEUR-WS.org. Retrieved from

Carriero, V. A., Gangemi, A., Mancinelli, M. L., Marinucci, L., Nuzzolese, A. G.,Presutti, V., & Veninata, C. (2019). *ArCo ontology network and LOD on italian cultural heritage.* In A. Poggi (Ed.), Proceedings of the first international workshop on open data and ontologies for cultural heritage, odoch@caise 2019 (Vol. 2375, pp. 97–102). CEUR-WS.org. Retrieved from

Chao, H., Delbridge, M., Kenderdine, S., Nicholson, L., & Shaw, J. (2018). Kapturing *Kung Fu: Future proofing the Hong Kong martial arts living archive*. In Digital echoes (pp. 249–264). Springer.

Groth, P., & Dumontier, M. (2020). *Introduction - FAIR data, systems and analysis*. Data Sci., 3 (1), 1–2. Retrieved from https://doi.org/10.3233/ doi: 10.3233/ds-200029

Hou, Y., Picca, D., Egloff, M., & Adamou, A. (2021). *Digitizing intangible cultural heritage embodied: state of the art.* Journal on Computing and Cultural Heritage. (To appear)

Picca, D., & Egloff, M. (2017). *DHTK: the digital humanities toolkit.* In A. Adamou, E. Daga, & L. Isaksen (Eds.), Proceedings of the second workshop on humanities in the semantic web (whise II) (Vol. 2014, pp. 81–86). CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-2014/

Sloetjes, H., & Seibert, O. (2016). *New facets of the multimedia annotation tool ELAN.* In M. Eder & J. Rybicki (Eds.), 11th annual international conference of the alliance of digital humanities organizations, DH 2016 (pp. 888–889). ADHO. Retrieved from http://dh2016.adho.org/abstracts/161

Notes

1. CROSSINGS knowledge graph sources, https://github.com/CROSSINGS/kg
2. DOI for data citation: https://doi.org/10.5281/zenodo.5886867
3. Hakka Kung Fu portal, http://www.hakkakungfu.com/

The case for DH in Literary scholarship

Pierazzo, Elena
elena.pierazzo@univ-tours.fr
University of Tours, France

Not all Humanities have been equally touched by the digital. For textual scholarship, history and linguistics, for instance, we can have a substantial number of scholarly contributions, particularly when we include experiences embodied in projects and resources. However, comparatively speaking, digital literary criticism has had few followers. An exception are Computational Literary Studies (CLS) that apply quantitative methods to large amount of literary and bibliometric data. Linked to the methods of distant reading [Moretti, 2005], this approach enjoys great success today, while web resources like Voyant, software like Gephi, and programming environments like R, have made text mining very accessible, even for those with limited computer skills. Linked to this approach, stylometry and authorship attribution are also thriving. Particularly mediatized researches are the initiatives that led to "unmasking" the identies of Robert Galbraith, a pseudonym of J.K. Rowling, and Elena Ferrante [Joula, 2015; Tuzzi and Cortelazzo, 2018]. However, literary criticism connected to close reading seems almost absent from the DH radar. The CATMA tool, designed to define personalized tagsets for (mainly) literary analysis [Meister 2020], represents a bright exception. Meister, in fact, is one of the few scholars that has engaged with digital literary criticism and digital hermeneutics; the latter has been explored also by Van Zundert (2016) and Ramsey (2011), but from a quantitative perspective. Relatively few scholars in DH have to addressed literary criticism with qualitative approaches, which are, conversely, among the most important for non-digital literary scholars.

The reasons for this absence are probably to be found in the controversies about the use of markup within texts that have inflamed the scholarly community since the Eighties. The act of adding explicit markers in the text has been subjected to scrutiny, as it is perceived (rightly) as a harbinger of interpretation and this fact has been (and is, to a certain extent, still) perceived as an invasion, a disfigurement of the text; Cummings (2008) gives a vivid account of the debate and reflects on how it has limited the use of TEI for literary criticism. The argument goes that once the text is marked up, it cannot be reused by others because the interpretation added by the encoder would make it unusable. According to this vision, digital texts must be made available in their most neutral and objective form, and any form of annotation, including editorial, must be avoided. Sperberg-McQueen 1991 and Cummings 2008, amongst others, have tried to address the issue, and I have argued elsewhere on the hermeneutic fallacy of the category of objectivity [Pierazzo 2015]; but these methods remain far from impacting "the Humanities at large" and in particular the literary scholars [Meister 2020]. However, in order to contextualize this debate, one should go back to when this controversy was born. The urgency of those years was to put texts online, to create literary corpora for concordances and the study of word frequencies; at the time, digital acquisition of texts, the transformation of the printed into sequences of characters to be analyzed by computers (Machine Readable Form) was mostly done by hand, with an enormous expenditure of time and energy. The emphasis was therefore on making texts available and on the need of not repeating work. Researchers did not want to work with texts full of manually added codes which then had to be removed just as manually in order to reuse the texts.

It is worth noting, though, how this discourse hides the concept of DH as a service: the goal was thought produce resources for others to do "real" research. This argument is not only dangerous, condemning DH to a mere service, but also wrong, as text, any text, can only be the result of the dialectical compromise between the source documents that contains it and scholars that interpret it (even when they "only" transcribe it), and therefore no text can ever be considered objectively neutral [Pierazzo, 2015]. Today conditions have changed: most literary texts are digitally available in many versions, not to mention the plethora of tools and methods to "get rid of" markup, therefore the objections do not stand in the same way.

Another obstacle for the uptake of DH in literary studies is the conviction that close reading and critical interpretation only require a reader, a text, and a (printed) essay, and therefore computers, in this context, are useful as typewriters [Kirschenbaum, 2016]. Yet, the lack of experimentation and engagement of the scholarly community in DH for literary analysis does not allow for a clear assessment of the epistemological added value of using computers for one or few texts at a time. But shouldn't be this the moment for rethinking Digital Literary Studies? Couldn't we at least try to use markup, ontologies and other methods to understand a text, or answer questions about interpretation?

The paper will present some experiences at the University of Tours using TEI markup for the history of

ideas, and ontologies and databases for analysis of fictional entities (people and places). We have applied these methods to works by Boccaccio, the Vite by Vasari, and to a small corpus of librettos of the 17th century. These experiments are showing promising results, not only in literary terms, but also on a largely methodological perspective, with colleagues and researchers finding themselves challenged and enticed by DH heuristics.

Conditions are ripe for experiences and discussions in order to evaluate the impact of DH in literary studies, particularly in the light of the advancements in HTR and other types of CLS that have the potentials of bringing a large amount of unknown and understudied texts into the literary arena. This could truly change our perspectives and understandings on literature, but we need to sharpen our hermeneutical tools first.

Bibliography

Cummings, J., 2008. The text encoding initiative and the study of literature. In *A companion to digital literary studies* (pp. 451-76), Blackwell.

Kirschenbaum, M.G., 2016. What is digital humanities and what's it doing in English departments?. In *Defining Digital Humanities* (pp. 211-220). Routledge.

Juola, P., 2015. The Rowling case: A proposed standard analytic protocol for authorship questions. *Digital Scholarship in the Humanities*, 30(1): 100-113.

Pierazzo, E., 2015. *Digital scholarly editing: Theories, models and methods*. Routledge.

Ramsay, S., 2011. *Reading Machines: Toward and Algorithmic Criticism*. University of Illinois Press.

Sperberg-McQueen, C.M., 1991. Text in the electronic age: Texual study and textual study and text encoding, with examples from medieval texts. *Literary and Linguistic Computing*, 6(1): 34-46.

Tuzzi, A. and Cortelazzo, M.A., 2018. What is Elena Ferrante? A comparative analysis of a secretive bestselling Italian writer. *Digital Scholarship in the Humanities*, 33(3): 685-702.

Van Zundert, J.J., 2016. Screwmeneutics and hermenumericals: the computationality of hermeneutics. *A companion to digital humanities*. (pp. 331-347) Blackwell.

Minor Labels: Detecting Genre in Pitchfork Reviews, a "Metamodularity" Network Analysis

Porter, J.D.

porterjd@sas.upenn.edu
Price Lab for Digital Humanities at the University of Pennsylvania, United States of America

Varner, Stewart

svarner@upenn.edu
Price Lab for Digital Humanities at the University of Pennsylvania, United States of America

Pitchfork.com has published music reviews, news, interviews, feature stories since 1995. Growing out of 1990s zine culture, *Pitchfork* took advantage of the affordances of the internet early and has since become the self-proclaimed "most trusted voice in music." In that time, they have reviewed over 23,000 albums by more than 10,000 artists. When it began, *Pitchfork* was known for its attention to alternative/punk/indie rock and, though it has increased its coverage of hip-hop, pop, and more mainstream music over the years, it remains a source of information about unconventional and emerging artists across genres.

Though they often cite an extraordinary number of genres, sub-genres, and sub-sub-genres in their reporting, they stick to a very conservative list of nine official genre tags for most of their reviews.

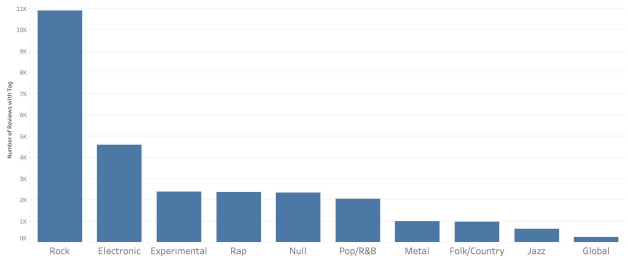


Figure 1:
Reviews by genre in Pitchfork. *Note that one review may have several genres, each of which would be counted in this graph. The "Null" category describes a small subset of reviews that had no genre tags.*

Debates over the function of genre are common in music as well as literature, film, and culture studies generally (Sanneh, 2021; Lena, 2012; Brackett, 2016). For an outlet like *Pitchfork*, they function, at least in part, to set expectations for readers. Yet, given the scale and the breadth of the types of music covered by *Pitchfork*, these nine highly unevenly distributed tags are somewhat uninformative. While the genre tags alone may do little to set expectations, the reviews themselves do this work in at least one highly characteristic way: by comparing the artist under review to other artists who share similar qualities. Fortunately for us, *Pitchfork* does this in an easily minable fashion by maintaining individual pages for several

thousand artists, and then using links to these pages when the artists are mentioned in reviews.

> The phrase "roots rock" feels especially irrelevant in New Orleans, the delta of the Mississippi River along where rock'n'roll, jazz, and the blues all supposedly began. At the dawn of the 1970s, the most forward-thinking musicians in the bayou weren't simplifying their sound. They were expanding it with traditions from their own communities: The Meters flirted with second line rhythms, while Dr. John dosed his swamp shamanism with Mardi Gras pomp. Rediscovering roots was nonsense—these artists had never parted with them.

Figure 2

Figure 2 is a screenshot from a review. Clicking on "The Meters" or "Dr. John", indicated by red underlining, will take you to the individual pages for each respective artist. We scraped all of the reviews and relevant metadata, including the artist links. With that data set, we created networks of artists where edges are drawn between any artists who have links in the same review.

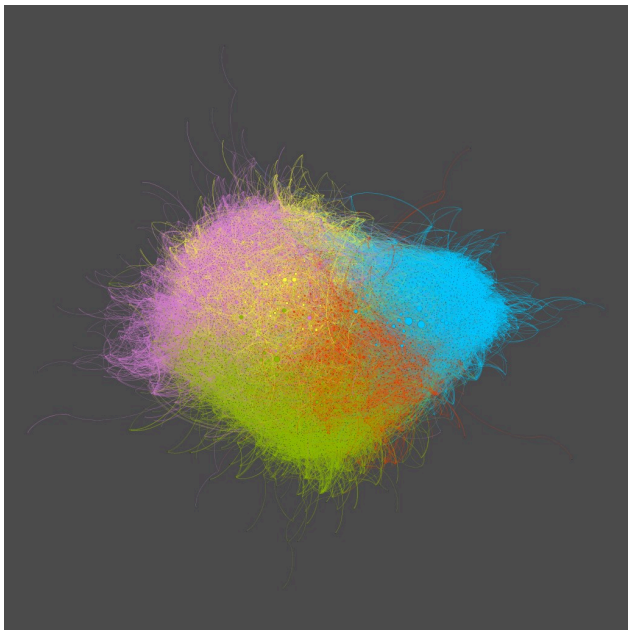


Figure 3.
Nodes are artists; edges reflect co-presence in reviews. Node size shows edge count. Color shows detected community.

Tools like Gephi can depict sub-groups in a network like the one we created via community detection (calculated in Gephi using the Louvain method (Blondel et al, 2008)). However, there are important limitations to this method.

Because it is non-deterministic, the precise membership of each group, and even the total number of groups, can vary each time the algorithm is run. To work around this, we introduce a method we call "metamodularity". We simply ran Louvain community detection on the artist network 10,000 times. [1] Though there are many other possibilities that would achieve similar ends, including examining the communities detected during the "passes" from which the Louvain method constructs its final groups, our approach has several key advantages: It is simple to run, easy to understand, and eschews a non-deterministic approximation in favor of data about the probability of particular groupings.

Using this method, we can show how often any two artists were sorted into the same group. For instance, the jazz musician Alice Coltrane was grouped with John Coltrane 10,000 times, with Sly and the Family Stone 4,939 times, and with Guns n' Roses one time. This gives us a more reliable and comprehensible picture of the level of connection between the artists.

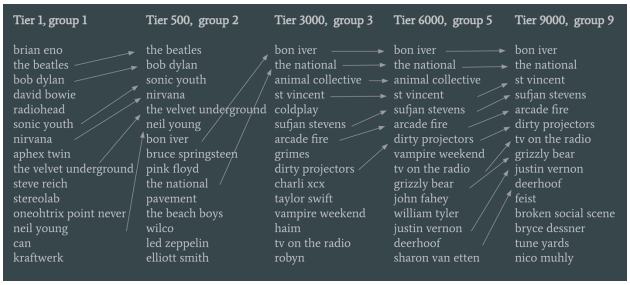


Figure 4:
Showing the top artists (sorted by number of total connections with other artists) who group with Bon Iver at increasing thresholds of connectedness

Using the results from this method, we examined every group of at least 25 artists at six different thresholds and renamed the groups to reflect our assessment of the underlying artistic community. For instance, we called the rightmost group in Figure 4 "00's Indie Rock".

It is worth underscoring that our group names are subjective. Nonetheless, they help to show the relationship between groups at different tiers. The Sankey diagram, Figure 5, makes clear the branching of closely-knit artist communities from larger, more loosely connected groups.

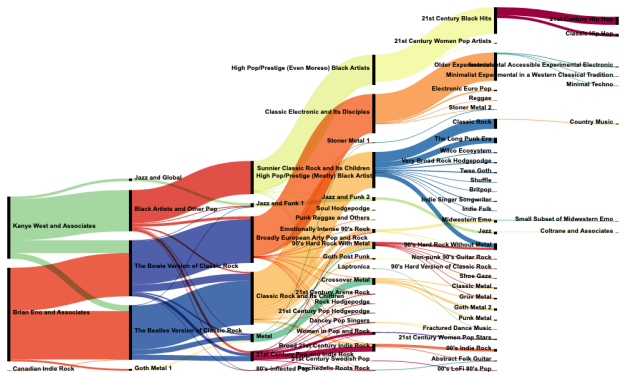


Figure 5

The richness of these results points to many interesting findings about the operation of genre in *Pitchfork*; we will mention just one here. Some of the genres suggested by Figure 5 are incredibly specific—e.g., our names for groupings of metal artists include classic, goth, punk, grüv, and crossover. One notable exception is a group of predominately African American musicians, which is remarkably large and stable up until the 9,000 threshold. The artists within it range from instrumental hip hop composer Flying Lotus to Motown legend Marvin Gaye to trap rapper Young Thug to R&B singer-songwriter Sade to funk innovators Funkadelic to crossover hip hop star Cardi B. This is a far more capacious group along aesthetic, market, and even historic grounds than we see in, e.g., the heavy metal clusters at the same metamodularity threshold. This may reflect real-world connections, since many artists in this group have worked together in various ways, perhaps more often than metal bands have. Or this may reflect a structural difference in the way that writers at *Pitchfork* have covered Black artists relative to their reviews of white musicians, particularly in the early years of the publication. In any case, it is a noteworthy difference in the shape of genre in this corpus.

This finding will be one of three concluding points in the talk. We will also discuss how, in this dataset, metamodularity based solely on artists' connections seems to demonstrate Carolyn Miller's description of genre as "social action" rather than some kind of top-down, taxonomic structure (Miller, 1984). We will also reflect on the potential use of metamodularity as a more broadly useful method for understanding (and depicting) network structures.

Bibliography

Blondel, V., Guillaume, J., Lambiotte, R. and Lefebvre, E. (2008). Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment*. 2008: 10. 9.

Brackett, D. (2016). *Categorizing Sound: Genre and 20th Century Popular Music*. Oakland: University of California Press.

Lena, J. (2012). *Banding Together: How Communities Create Genres in Popular Music*. Princeton: Princeton University Press.

Miller, C. (1984). Genre as Social Action. *Quarterly Journal of Speech*. Vol 70, 2; 151-76.

Sanneh, K. (2021). *Major Labels: A History of Popular Music in Seven Genres*. New York: Penguin Press.

Notes

1. We used the NetworkX and Community libraries in Python. Some artists had no links, because they were never mentioned in an article with anyone else. To avoid very small communities and improve the legibility of results, we also filtered out 34 artists not connected to the main network. This left us with 7,524 unique artists.

Digitizing Derrida's Concept of Dissemination: From Returntocinder.com to Databyss.org

Reeder, Jake
jreed03@gmail.com
Birkbeck College, United States of America

"To write is to produce a mark that is a kind of machine, legible in a scribe 's absence." Jacques Derrida, *Signature, Event, Context* (1982)

Dissemination is an activity or practice; it is the process of grafting text into other contexts—or, to put it simply, **spreading the word** (Derrida, 1982). In this sense, the act of annotating a text is already a sort of dissemination. Indeed, when beginning work on the website returntocinder.com, a published index of margin notes from the works of Jacques Derrida, the authors regarded annotation as "additive, useful, social, a means to collaborate with a text, and a 'meta conversation running in the margin'" (Kalir and Garcia, 2021: 7). Returntocinder.com and its companion application, Databyss, affirm the social usefulness of the margin note, believing that the publication of notes, ideas, and quotations

in a hypertext format can accelerate these conversations, producing new notes, ideas, quotations and digital projects.

Inspired by Ted Nelson's original theoretical work in hypertext and his writing application, Xanadu, the authors understood that "non-sequential writing" could benefit from developing software and thus algorithms that process text (Nelson, 1987). This processing, as Kalir and Garcia note, is an act of annotation in the computer science sense, meaning "labeling data—images, text, and audio—for the purposes of identifying, categorizing, and training machines and algorithms" (Ibid.). That is to say, generating metadata is a form of annotation. The supposed juxtaposition between that of the writer writing in the margins and that of the algorithm (programmer) categorizing, databasing and hyperlinking, is exactly what Derrida's dissemination combines into a single concept. Both writer and developer are performing the same action of annotation/dissemination because the act of margin notes is already an act of labeling, an act of creating grammar, searchability, and organization, while the act of automatic indexing, databasing and hyperlinking is already an act of grafting new texts in the (hypertextual) margins. The margin note is already a minimal program and the hypertext program is a margin note; together, they facilitate the practice of digital dissemination through annotation.

The Databyss Foundation (the non-profit company that has created returntocinder.com and databyss.org) began, as mentioned, as a project to build an index of several of Derrida's works and publish it in a way that was true to his concept of dissemination. The data was originally collected "manually" in a long word processor document, but because Derrida sees every written thought as a graft, ready to connect to another graft, it didn 't make sense to publish this index as a single web page or even as a set of linear pages. Instead, we wrote the document using a grammar that would allow us to process and publish each note separately. Each paragraph began with a short code indicating the book from which the note was taken, followed by a page number. Notes were grouped under bolded headings that indicated a motif common to the notes below. We then wrote a parser script that broke the linear document into separate database entries, linking each to a table containing all of the sources and to another table that contained the motif headings.

The result was returntocinder.com, an online database of margin notes (quasi-quotations) from several of Derrida 's works that is searchable and navigable by way of an index of keywords and sources. After its initial launch, we discovered that one could easily add other authors to the database structure we had designed. This was quite exciting, as it meant that the **carrier of ideas** could, in a sense, have a life separate from the ideas that originally inspired them. Since its initial publication five years ago, returntocinder.com has grown to include notes from texts by over 50 authors and receives hundreds of visitors every day.

As more people used the site, researchers approached us hoping to build a similar resource for their work. Rather than suggest that they write a long document using our grammar and run it through our parser, we set about on a second project: The Databyss App. This web-based application codifies the grammar we used to generate returntocinder.com into a word processor-like interface. The short codes indicating sources and the bold headings indicating motifs are recognized by the app as you write them and linked to useful index pages that show other notes with the same tag(s). These links or tags are browsable and searchable in a sidebar to the left of the editor. We also set out to create a workspace that caters to humanities workers (researchers and students), with additional features like the ability to pull PDF highlights and annotations straight into your workflow, and search book and journal catalogs to generate Zotero-like citations. Future modules will facilitate interoperability with other annotation and bibliography applications such as Zotero, Hypothes.is and Readwise. When users share some or all of their Databyss notes, all the links remain intact and the workspace interface (i.e all the search options in the sidebar) is fully functional for the reader, so there is no need for a separate parsing and publishing step. True to Derrida 's concept of the written mark, these Databyss notes become self-disseminating machines.

In our presentation, we will first cover the theoretical foundations of our work and situate it within the context of prior hypertext resources, such as David Kolb's "Socrates in the Labyrinth" and Eric Steinhart's "Fragments of the Dionysian Body." We will then demonstrate how the current features of the Databyss application can be used for humanities research and introduce some use cases already proposed or practiced on Databyss (a Critical Race Theory resource, a film studies curriculum, notes for an academic podcast, student reading assignments, and the distribution of a professor 's lecture notes). Finally, we will discuss the ideas generated by our present work for the future of the application and foundation.

Bibliography

Derrida, Jacques. (1982). Signature Event Context: *Margins in Philosophy*. Translated by Alan Bass. Chicago: University of Chicago Press.

Kalir, Remi H., and Antero Garcia. (2021). *Annotation*. Cambridge: MIT Press.

Nelson, Theodor H. (1987). *Computer lib: Dream machines*. Redmond: Tempus Books of Microsoft Press.

Locating a national collection through audience research

Rees, Gethin
dr.gethin.rees@gmail.com
British Library, United Kingdom

Additional authors: Alex Hunt, National Trust; Valeria Vitale, Alan Turing institute; John Horgan, STRAT7 ResearchBods; Peter Strachan, STRAT7 ResearchBods.

This presentation explores the role of geography in public engagement with digital cultural heritage collections. It draws on audience research that examined public values and motivations alongside the use of location-based interfaces such as web maps. The research, conducted as part of the Locating a National Collection project (LaNC), is part of the Towards a National Collection programme and funded by the UK's Arts and Humanities Research Council. The programme aims to 'break down the barriers that exist between the UK's outstanding cultural heritage collections, with the aim of opening them up to new research opportunities and encouraging the public to explore them in new ways.'

Digital cultural heritage records are connected with geographical locations in diverse ways. Heritage sites are managed for the purposes of tourism, education and preservation by historic environment institutions (e.g. LaNC partners Historic England or Historic Environment Scotland). The objects, documents, and other records that constitute the collections of galleries, libraries, archives and museums (GLAMs) (e.g. LaNC partner British Library) are connected to the locations where they were made and used or those they depict and describe. Historic environment organisations use institutional systems such as Geographical Information Systems (GIS) that are based on coordinate data. Locations do not play this structuring role in the institutional systems of GLAMs where historical toponyms predominate. Digital humanities research has examined the implications of toponyms for visualisation (Gregory and Hardie, 2011). Our project seeks to derive value for a range of audiences by building bridges between GLAMs and Historic Environment collections using location.

Audience research method

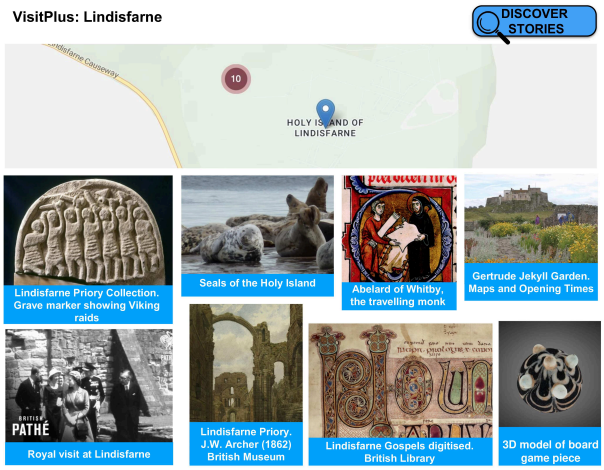
LaNC has used an innovative audience research methodology to understand how geography can help the public to engage with cultural heritage. Audience research was led by Alex Hunt at the National Trust working with Research Bods, a market research company and project researcher Valeria Vitale. The research examined three areas:

- Attitudes and behaviour around cultural heritage
- Values and their relations to place and geography
- Use of digital technologies in cultural heritage and beyond

It divided into two interlinked phases: a web survey followed by focus group interviews, both based on representative samples of UK public. This presentation focuses on the latter. After opening discussions, focus groups explored how location-based technologies might present heritage on the web drawing on simple 'powerpoint-slide' sketches of future interfaces, known as pretotypes (Savoia, 2011).

Visitplus

The first pretotype, 'VisitPlus', presented an interactive web map designed to enhance heritage visits. The pretotype offered the example of Lindisfarne (castle managed by National Trust and priory managed by English Heritage). The user can access related resources including visitor information alongside GLAM objects including Lindisfarne Gospels served in iiif by the British Library, paintings and 3D models from British Museum. Underlying VisitPlus is a gazetteer of heritage visitor locations numbering in the low thousands and typically defined as nationally or internationally significant (Timothy, 2014). Each location acts as a nexus, an entry point to access many collection items. The pretotype was designed to probe audience attitudes to aligning GLAM records with visitor sites. For example, are Lindisfarne visitors interested in viewing the gospels when they return? Focus-group participants perceived strong benefits in using VisitPlus and valued access to diverse and high quality resources in a single location. Although many saw potential uses to enhance understanding before and after visits others were confused about whether the intended use was exclusively on-site. Differences in historical veracity between linked resources and guidebooks were especially concerning for participants with a high interest in heritage. The 'VisitPlus' branding was somewhat problematic: the name implied a particular usage obscuring broader uses and value.



Thanks to Valeria Vitale for image.

Heritage for all

The ‘Heritage for All’ pretotype allowed users to explore the collections of GLAMs and historic environment organisations connected to the ‘place where they live now’. Addresses or coordinates referenced tens of thousands of precise locations like buildings or parks. Links to institutional web pages were accessed by clicking on locations. An example web map showed a London borough populated with text previews of content such as findspots of ancient coins, extant and disappeared historical buildings and literary references from famous novels, each with their own location. Much of this heritage has been typically defined as locally significant (Goodchild and Hill, 2008; Timothy, 2014, 34). Participants saw value in connections between varied collections such as literary references alongside the physical locations to which they refer. The visualisation of precise locations within a model akin to GIS connected cultural heritage to the extant physical environment, inspiring and informing in-person visits. For some, the diversity was confusing, with interactivity, curated themes or narratives favoured. Participants cited repeated usage as a limitation: why return after exploring the place where they live now and other familiar places? Yet the overriding response was that users found diverse collections connected to familiar locations hugely compelling.
The passionate feedback ‘Heritage for All’ evoked was influenced by renewed interest in local exploration resulting from travel restrictions due to the Covid-19 pandemic.



Thanks to Valeria Vitale for image.

Conclusion

Heritage significance has been understood at varied geographical scales such as the supra-national (e.g. UNESCO), national and sub-national, the latter encompassing regional or local scales (Ashworth and Larkham, 2013). Our audience research has shown that collections traditionally defined as locally significant often hold the most meaning for the public. Geography and community are intimately connected. Although outlining pathways to community impacts such as social cohesion is beyond the scope here, understanding how the public engage with geographical information is a critical step in this endeavour. A user-centred approach to web-map design that draws on digital-humanities visualisation research can help users to discover parts of collections that are significant to them (Roth, 2019). LaNC will go on to build a web map interface that tests these pretotype ideas seeking to understand how cultural heritage defined as locally significant might cohere as a national collection, as our relationships with geography are redefined in a post-covid world.

Bibliography

Ashworth, G, and Larkham, P. (2013). *Building A New Heritage (RLE Tourism)*. London: Routledge.

Gregory, I.N., and Hardie, A. (2011). Visual GISting: Bringing Together Corpus Linguistics and Geographical Information Systems. *Literary and Linguistic Computing,* 26 (3): 297–314. https://doi.org/10.1093/llc/fqr022

Goodchild, M. F., and Hill, L. L. (2008). Introduction to Digital Gazetteer Research. *International Journal of*

Geographical Information Science, 22 (10): 1039–44. https://doi.org/10.1080/13658810701850497

Roth, R. (2019). How Do User-Centered Design Studies Contribute to Cartography? *GEOGRAFIE,* 124 (2): 133–61. https://doi.org/10.37040/geografie2019124020133

Savoia, A. (2011). Pretotype It. https://drive.google.com/file/d/0B0QztbuDlKs_NzBjYWNiOGQtNmQyNi00OWE2LWI2YzktN2Y3YTEzM2VjYTNj/view?resourcekey=0-zeFx0Fesf3X3pRjYzIS1ZA (accessed 21 April 2022).

Timothy, D. J. (2014). Contemporary Cultural Heritage and Tourism: Development Issues and Emerging Trends. *Public Archaeology,* 13 (1–3): 30–47. https://doi.org/10.1179/1465518714Z.00000000052

Telescopic reading: Synthesizing meaning from reading at different scales

Ringler, Hannah
hringler@andrew.cmu.edu
Carnegie Mellon University, United States of America;
Illinois Institute of Technology, United States of America

Argamon, Shlomo
argamon@iit.edu
Illinois Institute of Technology, United States of America

Historically, computational humanities work has conceptualized a divide between close and distant reading. The humanities have a long history of close, detailed reading of individual texts. In 2000 though, Moretti formalized the concept of "distant reading," or looking at entire swaths of cultural or historical literatures without a focus on individual texts, which has become closely linked to corpus methodologies and has proven to be a useful framework for certain types of research questions.

Often though, close and distant reading can complement each other in the same analysis to open up a broader range of questions. Reading distantly, especially with computational tools, can reveal textual patterns that assist in developing and answering certain broader social-historical questions. To use these textual patterns to understand the texts themselves more deeply, though, we must turn back to individual texts to understand what the data mean. For example, in Grant et al.'s (2021) analysis of archival documents on global policies on refugees, they explain how the right to movement was a key part of arguments about Ugandan Asian resettlement, and suggest a historian should closely read these documents for a more in-depth interpretation. In this study and others like it, word lists of frequencies are only fruitfully transformed into interpretations by returning back to reading individual texts from the corpus.

But moving between distant and close reads of a corpus is both practically and theoretically difficult. When staring at a table or plot of word frequencies, where do you begin in trying to ascertain why certain words are common (or not) in a particular corpus and what that means for the corpus as a whole? Literary and rhetorical studies would remind us that there are many possible ways to interpret texts, and a "good" one is one that can be argued for well. The interpretive processes involved are not straightforward, neutral, or objective, even in distant reading, despite the seemingly-objective feel of data in computational work. But our interpretive processes for texts are largely based around reading one text at a time. How can we do this sort of reading with multiple texts or a whole corpus? In other words, how to fruitfully interpret computational data when it requires both close and distant reading, and how do we know it is done well? Several scholars have stressed the importance of iteration as part of the answer to this process (e.g., Rockwell and Sinclair, 2016; Guldi, 2018), but the practicalities of iterating with close and distant reading are complex and not straightforwardly combined into one cohesive, productive process.

In this project, we offer a methodological framework for interpreting computational models of texts, which we call multifocal reading. We sketch this framework, and illustrate it with a case study to demonstrate how to read strategically and in a well-justified way so as to gain deeper insight into corpora. In particular, we draw on past theoretical work on this problem of interpreting computational models of texts (Piper, 2015; Rockwell and Sinclair, 2016; Ringler, 2021) and move theory into practice by proposing a methodological framework in the form of a six-step process for interpreting models of corpora toward insight into the texts themselves. The steps are as follows:

1. Choose a body of texts
2. Create a widely-focused view
3. Form initial explanations
4. Narrow focus reading
5. Refine and synthesize explanations
6. Argue for your understanding

We focus on the development of hypotheses from computational results and demonstrate how traditional close and distant reading must be slightly modified into what we call wide focus and narrow focus reading, illustrating how narrow focus reading of specific texts can help the analyst probe and refine those hypotheses toward understandings that open up interpretive possibilities for large corpora. In other words, we argue that a text analytic hypothesis testing method does not close off interpretive possibilities

by merely attempting to prove or disprove certain textual interpretations, but rather contributes to the exploration of new and complex interpretative possibilities.

As a demonstration of our framework, we start with Underwood and Sellers' (2016) study on literary prestige. This study found that a logistic regression model distinguished prestigious from random 19th-century poetry with some accuracy and used this to talk about the "long arc of prestige" as imagined in literary history. We begin with their logistic regression model to ask, **what do we learn about prestigious 19th-century poetry?** A detailed treatment of this question was outside of the scope of the original study, but the study itself allows the question to be asked.

We use the logistic regression model, with past research on Victorian poetry, to develop initial hypotheses about prestige in poetry. We then probe these hypotheses through strategic high resolution reading, using theories of pathos and affect to guide our analysis. We consider these readings strategic in that we carefully choose sets of texts to read ranking high, medium, and low on various model features to systematically probe aspects of our hypotheses. Through this process, we find that not only is prestigious poetry more negative in tone (the result of the original study), but that prestigious poetry overall tends to focus more on creating a mood or feeling through the text, which is often dark, mysterious, and haunting. This finding expands theoretical understandings of Victorian poetry, as well as provides new insights and questions on how particular literary effects were achieved through recurring and specific linguistic forms.

Ultimately, this project offers a new way forward in interpreting computational data in the humanities, demonstrating how we can gain defensible, robust insights into corpora of texts through strategic reading in a way that opens up (rather than closes off) interpretive insights. By theorizing these interpretive processes, we hope to move towards addressing the perpetual "so what" and "we can't read it all" issues of distant reading; and by making these processes explicit, we hope to further clarify sometimes-opaque text analytic methods to make them more accessible to a diverse range of scholars.

Bibliography

Grant, P., Sebastian, R., Allassonnière-Tang, M., & Cosemans, S. (2021). Topic modelling on archive documents from the 1970s: global policies on refugees. *Digital Scholarship in the Humanities*, 36(4): 886-904.

Guldi, J. (2018). Critical search: A procedure for guided reading in large-scale textual corpora. *Journal of Cultural Analytics*, 3(1).

Moretti, F. (2000). Conjectures on world literature. *New Left Review*, 1: 54-69.

Piper, A. (2015). Novel devotions: Conversional reading, computational modeling, and the modern novel. *New Literary History*, 46(1): 63-98.

Ringler, H. (2021). 'We can't read it all': Theorizing a hermeneutics for large-scale data in the humanities. *Digital Scholarship in the Humanities*.

Rockwell, G., and Sinclair, S. (2016). *Hermeneutica: Computer-Assisted Interpretation in the Humanities.* Cambridge: The MIT Press.

Underwood, T., & Sellers, J. (2016). The Longue Duree of Literary Prestige. *Modern Language Quarterly*, 77(3): 321-344.

Archiviz: A Tool for the Interactive, Visual Exploration of Digital Archives

Rittenhouse, Brad
bcrittenhouse@gatech.edu
Georgia Institute of Technology, United States of America

Michney, Todd
todd.michney@hsoc.gatech.edu
Georgia Institute of Technology, United States of America

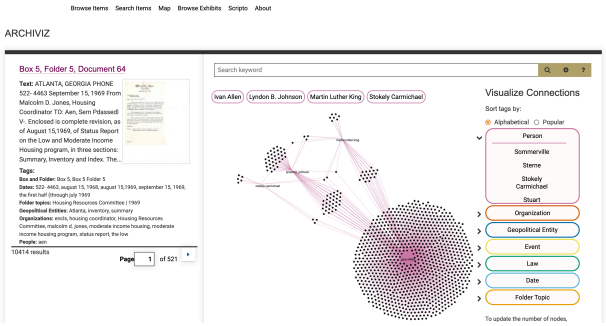
Acosta, Ines
iacosta6@gatech.edu
Georgia Institute of Technology, United States of America

In traditional, paper-based archives, finding aids afford basic means for discovery, but tend to return results shaped by researchers' preexisting knowledge and queries. With the digitization of archives, new analysis and visualization methods allow researchers to textually map large corpora and not only pinpoint specific materials, but also open new ways of navigation. [i] However, effective use of computational tools including natural language processing (NLP), named entity recognition (NER), and knowledge graphing often requires considerable technical sophistication, forming an access barrier for many researchers. [ii]

To help bridge this technical gap, we have been developing a toolset, currently funded by an NEH-ODH Digital Humanities Advancement Grant, that integrates large-scale text processing and data visualization capabilities into the open-source Omeka [iii] content management platform. With the tool, users can upload batches of documents, perform NER on them, and

manipulate an interactive graph that displays extracted entities (people, places, organizations, etc.) from the collection and how they interconnect in its component documents. The toolset's components [iv] were all specifically developed for (and tested by) non-technical researchers and community advocates, with all computational work taking place in a simple graphical user interface.

Our *Archiviz* toolset produces social network-style visualizations that connect results on the basis of the important people, places, organizations, and other entities mentioned. By applying NER to a collection, users can see the full breadth of their research subject at a glance, and explore connections they may not have known exist. [v] What our implementation offers that is new, is the efficient display of interrelationships between large numbers of nodes (named entities) combined with the possibility of rapid navigation to search results (documents). In addition, the interface moves beyond traditional query-based archival searching, "flattening" out a collection to show results beyond a researcher's interests or knowledge.



We are continuing to refine *Archiviz* in ways intended to be intuitive, usable, and readily adoptable by users from a variety of different backgrounds. A primary design goal of the project has been to make it accessible not just to relatively tech-savvy academics, but more importantly, communities beyond the academy. It is our hope that the tool may particularly benefit disadvantaged communities—which in the United States have often faced pressure from gentrification and urban redevelopment, with consequent coercive displacement from historical neighborhoods. Residents in such areas often lack the infrastructure to collect, preserve, and interpret local history. As such, we have partnered with various community groups and activists throughout the process to ensure that the tool can be useful to them. Most ambitiously, in 2019 we hosted a "Community Researcher Workshop" for Atlanta-based librarians, archivists, community organizers, nonprofit staffers, and students to explore the Mayor Ivan Allen Digital Archive that has served as our first test corpus, using a prototype of our toolset. [vi]

User testing of the interface during this workshop produced very positive feedback, with participants calling the platform "incredibly useful," with the "potential to break down traditional barriers of why people are hesitant to use archives." A GLAM (Gallery, Library, Archive, Museum) researcher noted that it "is something almost any archive or library could utilize to their advantage," while a community advocate stated that she "wanted the information for myself, but also…to share with my fellow residents." With an additional two years of grant-funded development since this workshop, we are excited to share our work with the DH community. While we have integrated much of the feedback from the workshop, future plans include capabilities to process a wider spectrum of digital, textualized media—documents, video and audio converted to text, and even images identified and categorized with computer vision making our tool relevant for a more diverse array of communities and collections. [vii]

Notes

i. See Graham, S., Milligan, I., and Weingart, S. (2015), *Exploring Big Historical Data: The Historian's Macroscope*, Imperial College Press, London; Morrissey, R. (2015), "Archives of connection: 'Whole network' analysis and social history," *Historical Methods*, vol. 48, no. 2, pp. 67–79; Duff, W., and Haskell, J. (2015), "New uses for old records: A rhizomatic approach to archival access," *American Archivist*, vol. 78, no. 1, pp. 38-58; Putnam, L. (2016), "The transnational and the text-searchable: Digitized sources and the shadows they cast," *American Historical Review*, vol. 121, no. 2, pp. 377–402; Edelstein, D., Findlen, P., Ceserani, G., Winterer, C., and Coleman, N. (2017), "Historical research in a digital age: Reflections from the Mapping the Republic of Letters project," *American Historical Review*, vol. 122, no. 2, pp. 400–424; and Hoekstra, R. and Koolen, M. (2018), "Data scopes for digital history research," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, vol. 52, no. 2, pp. 79–94.

ii. Piotrowski, M. (2012), *Natural Language Processing for Historical Texts*, Synthesis Lectures on Human Language Technologies, vol. 17, Morgan & Claypool Publishers, San Rafael; Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., and Gómez-Berbís, J.M. (2013), "Named Entity Recognition: fallacies, challenges and opportunities," *Computer Standards and Interfaces*, vol. 35, no. 5, pp. 482–489; Shahin, S. (2016), "When scale meets depth: Integrating Natural Language

Processing and textual analysis for studying digital corpora," *Communication Methods and Measures*, vol. 10, no. 1, pp. 28–50; Srinivasa-Desikan, B. (2018), *Natural Language Processing and Computational Linguistics: A Practical Guide to Text Analysis with Python, Gensim, spaCy, and Keras*, Packt Publishing, Birmingham and Mumbai; Ehrmann, M., Romanello, M., Flückiger, A. and Clematide, S. (2020), "Named Entity Recognition and linking on historical newspapers," in Arampatzis, A. et al. (eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Springer International, Cham, pp. 288–310.

iii. We developed *Archiviz* for the Omeka Classic version; technical and user support for our plugin will be available in February 2023 in the Omeka plugin library. Additional availability is planned through github and a Docker container for ease of use.

iv. The primary tools and plug-ins used to construct Archiviz include Harvard's Elasticsearch Omeka plugin for enhanced search capabilities, Google Tesseract for OCR, spaCy 's NLP functionality, particularly its NER functionality, with current development focusing on the integration of the Science Museum Group's Heritage Connector for better entity identification, linking, and disambiguation. The graphing of this information is primarily performed with the force-directed graphing of d3.js.

v. For a similar application, see Tumbe, C. (2019), "Corpus linguistics, newspaper archives and historical research methods," *Journal of Management History*, vol. 25, no. 4, pp. 533–549.

vi. On the issues here, see Flinn, A., Stevens, M., and Shepherd, E. (2009), "Whose memories, whose archives? Independent community archives, autonomy and the mainstream," *Archival Science*, vol. 9, no. 71, pp. 71–86.

vii. Wiriyathammabhum, P., Summers-Stay, D., Fermüller, C., and Aloimonos, Y. (2017), "Computer vision and Natural Language Processing: Recent approaches in multimedia and robotics," *ACM Computing Surveys*, vol. 49, no. 4, pp. 1–44.

Web Services for Voyant: LINCS, Voyant and NSSI

LINCS, Voyant and NSSI

Rockwell, Geoffrey Martin
grockwel@ualberta.ca
University of Alberta, Canada

Hervieux, Natalie
nhervieu@ualberta.ca
University of Alberta, Canada

Zafar, Huma
hzafar@uottawa.ca
University of Ottawa, Canada

Land, Kaylin
kaylin.land@mail.mcgill.ca
McGill University, Canada

MacDonald, Andrew
andrewjames.code@gmail.com
McGill University, Canada

Barbosa, Denilson
denilson@ualberta.ca
University of Alberta, Canada

Frizzera, Luciano
lucaju@me.com
Concordia University, Canada

Ilovan, Mihaela
ilovan@ualberta.ca
University of Alberta, Canada

Brown, Susan
sbrown@uoguelph.ca
University of Guelph, Canada

Introduction

It is difficult to identify named entities like people and places in long texts and even more difficult to connect the entities that you find to the rich network of information available on the web. In this paper we describe work supported by the LINCS (Linked Infrastructure for Networked Cultural Scholarship) project to make named entity recognition available to scholars through Voyant and its extension Spyral. In this talk we will:

First, describe the development of NSSI, a set of named entity recognition (NER) tools that are also available as web services for other tools like Voyant to use.

Second, describe how Voyant can use NSSI as a web service to process a text by adding named entity recognition.

Third, describe how Spyral, the notebook programming extension of Voyant, can be used for more sophisticated control of the process of named entity recognition, extraction, and use in Voyant.

Finally, we will conclude by discussing how NSSI and Spyral will be linked into the LINCS infrastructure to allow scholars to connect their enriched data to that of others.

Background on LINCS

Humanists tend to be interested in named people, named places and particular organizations over time. NER tools let humanists identify mentions in text referring to the people, places, organizations and other entities discussed in large collections without having to manually comb through them. Good tools like the Stanford Named Entity Recognizer (Finkel et al. 2005) have been available for some time, but are difficult to use if you are not familiar with command line tools and not connected with other resources.

The LINCS project, led by Susan Brown at the University of Guelph, is funded by the Canadian Foundation for Innovation to develop shared infrastructure for linked open data. To that end LINCS is working with teams at the University of Alberta and McGill University to develop new NER tools and to connect them to easy-to-use text analysis environments like Voyant.

NSSI

NSSI, or NERVE Secure Scalable Infrastructure, is an application that bundles natural language processing tools, making them simple to use and combine into workflows common to the digital humanities (Zafar 2021). This framework was developed as part of the LINCS project, with the intent to decouple the backend NER tools from the existing Named Entity Recognition Vetting Environment (NERVE) user interface developed by the Canadian Writing Research Collaboratory. This separation allows us to continue using those NER services for NERVE, while making them accessible to other tools such as Voyant and Spyral.

NSSI's design focuses on modularity, with each tool connected as a service that can be used individually or within a larger set of steps. For NER in particular, we have integrated Stanford NER which otherwise requires programming knowledge to use, since it does not come with its own API. With NSSI, a tool such as Spyral can make an API call that includes input text or XML and retrieve the named entities when processing completes. In the presentation we will briefly describe the NSSI infrastructure.

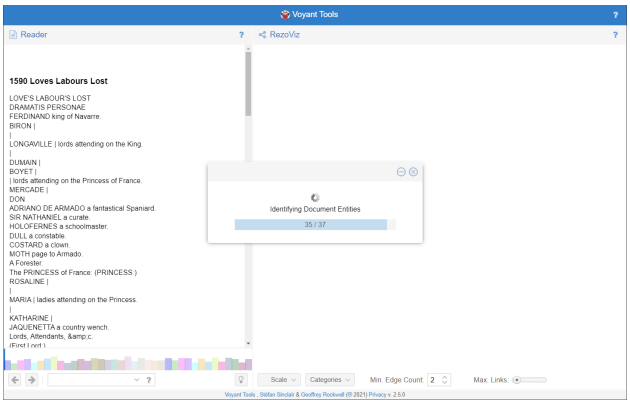


Figure 1:
Experimental RezoViz NER Interface in Voyant

Voyant and Spyral

Voyant Tools is a suite of text analysis and visualization tools that are widely used with over 100,000 users in the last six months. The tools are available in the browser so they don't need to be installed, though you can download them and run them locally (Rockwell & Sinclair 2016). In the presentation we will show how Voyant can call the NER tools in NSSI and display the found entities as a list for further use. We will also describe the usability testing conducted on ResoViz through the LINCS project.

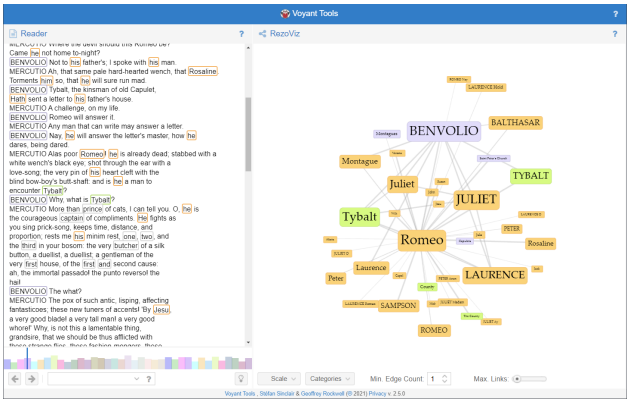


Figure 2:
ResoViz Social Network Visualization

Voyant is also being extended with a notebook programming environment called Spyral (Land et al. 2021; Rockwell et al. 2021). Spyral is, like Observable, an in-browser notebook programming environment that uses JavaScript as the programming language. The difference between Spyral and other notebook environments like Mathematica or Google Colab is that a) the notebooks are maintained on the server so that, again, there is no installation needed and b) Spyral is an extension to Voyant.

This means that you can save what you see in Voyant as a notebook with an interactive panel of results embedded in the notebook. Then you can document your results, add more interactive panels, and process the results. In the presentation we will show how Spyral can be used to extend the work with NSSI possible with Voyant and to edit and document results.

Next Steps

The paper will conclude by describing the next steps in the larger project, and those are to allow users to connect named entities in their texts to other data about the entities available through the LINCS triple store and other open data resources like Wikidata (Vrandečić 2012). The ultimate goal is to provide scholars with linked infrastructure where data about entities like people or novels can be annotated and connected with that of other projects.

Links

Google Colaboratory (Colab): https://colab.research.google.com/

LINCS project: https://lincsproject.ca/

Stanford Named Entity Recognizer: https://nlp.stanford.edu/software/CRF-NER.html

Voyant Tools: https://voyant-tools.org and Spyral: https://voyant-tools.org/spyral

Bibliography

Finkel, J. R., Grenager, T., and Manning C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf (accessed 21 May 2022).

Zafar, H. (2021). Linked Data Conversion using Microservices [video file]. Zenodo. https://doi.org/10.5281/zenodo.6551465 (accessed 21 May 2022).

Land, K., MacDonald, A. and Rockwell, G. (2021). Spyral Notebooks as a Supplement to Voyant Tools. CSDH-SCHN 2021 conference online. http://dx.doi.org/10.17613/2bsr-xp53 (accessed 21 May 2022).

Rockwell, G. and Sinclair, S. (2016). Hermeneutica: Computer-Assisted Interpretation in the Humanities. Cambridge, Massachusetts, MIT Press.

Rockwell, G., Land, K., and MacDonald, A. (2021). Social Analytics Through Spyral. Pop! Public. Open. Participatory. no. 3 (2021-10-31). https://popjournal.ca/issue03/rockwell (accessed 21 May 2022).

Vrandečić, D. (2012). Wikidata: A new platform for collaborative data collection. In Proceedings of the 21st international conference on world wide web, pp. 1063-1064.

From *Cyclopaedia* to *Encyclopédie*: Using Machine Translation and Sequence Alignment to Identify Encyclopedia Articles across Languages

Roe, Glenn
glenn.roe@sorbonne-universite.fr
Sorbonne University, France

Olsen, Mark
MarkyMaypo57@gmail.com
ARTFL Project, University of Chicago

Morrissey, Robert
rmorriss@uchicago.edu
ARTFL Project, University of Chicago

It is well known that the great 18th-century French *Encyclopédie* began first as a modest translation project of Ephraim Chambers' *Cyclopaedia* in 1745. And, although their project grew into something much more significant, the *Enyclopédie* editors (Diderot and d'Alembert) were not shy in incorporating translations of the *Cyclopaedia* as filler for their expanded work. Indeed, as Paolo Quintili remarks, 'the they left a good part of these articles almost unchanged, or with only minor changes' (Quintili, 1996: 75). Given the scale of the two works under consideration, however, systematic evaluation of the extent of the *philosophes'* use of Chambers has remained, even today, a daunting task. John Lough, in 1980, framed the problem thusly: 'So far no one has had the patience to make a detailed study of the exact relationship between the text of Diderot's *Encyclopédie* and the work of Ephraim Chambers. This would no doubt require several years of arduous toil devoted to comparing the two works article by article' (Lough, 1980: 221).

Recent developments in machine translation and sequence alignment now offer new possibilities for the systematic comparison of digital texts across languages. This paper outlines some recent experimental work in leveraging these new techniques in an effort to reduce the

'arduous toil' of textual comparison through automatic translation. In essence, we aimed to generate French translations of *Cyclopaedia* articles and then use sequence alignment to identify similar passages also found in the *Encyclopédie* [1].

We examined two of the most widely-used resources in this domain, Googe Translate and DeepL. Both systems provide useful APIs as part of their respective subscription services, and both provide translations based on cutting-edge neural network language models. While DeepL provided somewhat more satisfying translations from a reader's perspective, we ultimately opted to use Google Translate for the ease of its API and its ability to parse TEI-XML. The latter is of critical importance as we wanted to keep the overall document structure of our dictionaries to allow for easy navigation between the versions.

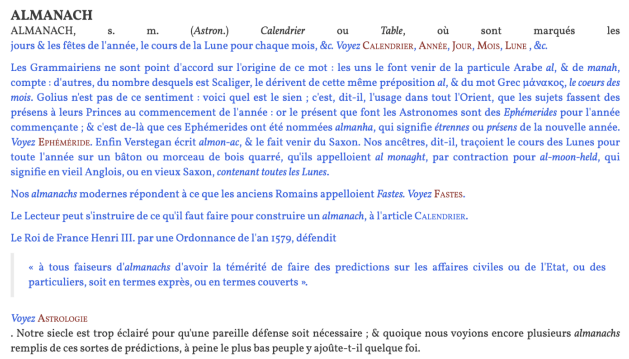
Our objective here was *not* to produce a good translation of the text, or even one that might serve as the basis for a readable edition. Rather, this machine-generated edition serves as a 'pivot-text' between the two corpora, allowing for an automatic comparison of the two (or three) versions using ARTFL's highly fault-tolerant sequence alignment package, Text-PAIR [2]. In order to determine the parameters for this task, we ran a series of tests with different matching parameters on a representative selection of 100 articles where Chambers was identified as the possible source. It is important to note that even with the best parameters, which we adjusted to get favourable recall and precision results, we were only able to identify 81 of these 100 articles.

Once settled on the optimal parameters, we then used Text-PAIR to generate both an alignment database, for interactive examination, and a set of static results tables. The alignment database contains some 7,304 aligned passage pairs. The system allows queries on metadata, such as author and article title as well as words or phrases found in the aligned passages. Each aligned passage is presented as a facing page representation and the user can toggle a display of the variations between the two aligned passages. As seen below, the variations between the texts can be extensive (*fig. 1*).



Figure 1.
Text-PAIR interface showing differences in the article "Air".

Text-PAIR also contextualises results back to the original document(s). For example, the following is the article "Almanach" by d'Alembert, showing the aligned passage from Chambers in blue (*fig. 2*).


ALMANACH
ALMANACH, s. m. (*Astron.*) *Calendrier* ou *Table*, où sont marqués les jours & les fêtes de l'année, le cours de la Lune pour chaque mois, &c. *Voyez* Calendrier, Année, Jour, Mois, Lune , &c.

Les Grammairiens ne sont point d'accord sur l'origine de ce mot : les uns le font venir de la particule Arabe *al*, & de *manah*, compte : d'autres, du nombre desquels est Scaliger, le dérivent de cette même préposition *al*, & du mot Grec μάνακος, *le coeurs des mois*. Golius n'est pas de ce sentiment : voici quel est le sien ; c'est, dit-il, l'usage dans tout l'Orient, que les sujets fassent des présens à leurs Princes au commencement de l'année : or le présent que font les Astronomes sont des *Ephémerides* pour l'année commençante ; & c'est de-là que ces Ephémerides ont été nommées *almanha*, qui signifie *étrennes* ou *présens* de la nouvelle année. *Voyez* Ephéméride. Enfin Verstegan écrit *almon-ac*, & le fait venir du Saxon. Nos ancêtres, dit-il, traçoient le cours des Lunes pour toute l'année sur un bâton ou morceau de bois quarré, qu'ils appelloient *al monaght*, par contraction pour *al-moon-held*, qui signifie en vieil Anglois, ou en vieux Saxon, *contenant toutes les Lunes*.

Nos *almanachs* modernes répondent à ce que les anciens Romains appelloient *Fastes*. *Voyez* Fastes.

Le Lecteur peut s'instruire de ce qu'il faut faire pour construire un *almanach*, à l'article Calendrier.

Le Roi de France Henri III. par une Ordonnance de l'an 1579, défendit

« à tous faiseurs d'*almanachs* d'avoir la témérité de faire des predictions sur les affaires civiles ou de l'Etat, ou des particuliers, soit en termes exprès, ou en termes couverts ».

Voyez Astrologie
. Notre siecle est trop éclairé pour qu'une pareille défense soit nécessaire ; & quoique nous voyions encore plusieurs *almanachs* remplis de ces sortes de prédictions, à peine le plus bas peuple y ajoûte-t-il quelque foi.

Figure 2.
Article "Almanach" with shared Chambers passages in blue.

In this instance, d'Alembert reused almost all of Chambers' original article "Almanac", with some minor variations, but does not to appear to have indicated the source of the first part of his article.

To accumulate results and to refine their assessment, we developed an evaluation algorithm for each alignment, with parameters based on the length of the matching passages and the degree to which the headwords were close matches. This simple evaluation model eliminated a significant number of false positives, which we found were typically short text matches between articles with different headwords. The output of this algorithm resulted in two tables, one for matches that were likely to be valid and one that was less likely to be valid, based on these simple heuristics.

In all, we found some 3,778 articles in the *Encyclopédie* that upon evaluation seem highly similar in both content and structure to articles in the 1741 edition of Chambers' *Cyclopaedia*. Whether or not these articles constitute real acts of historical translation is the subject for another, or several other, articles. There are simply too many outside factors at play, even in this rather straightforward comparison, to make blanket conclusions about the editorial practices of the *encyclopédistes* based on this limited experiment [3]. What we can say, however, is that of the 1,081 articles that include a 'Chambers' reference in the *Encyclopédie*, we only found 689 with at least one matching passage, although even here, the recall may in fact be higher than the numbers suggest, given that some citations function more like cross-references. Nonetheless, beyond testing this ground truth, we are also left with the rather astounding fact of 3,089 articles with no reference to Chambers whatsoever, all of which seem to be at least somewhat related to their English predecessor.

Notes:

[1] Our two comparison datasets are the ARTFL *Encyclopédie* and the recently digitised ARTFL edition of the 1741 Chambers *Cyclopaedia*. See https://artflsrv03.uchicago.edu/philologic4/encyclopedie1117/ and https://artflsrv03.uchicago.edu/philologic4/chambers_new/. The 1741 edition was selected as it was one of the likely sources for the translation original project and we were able to work from high quality pages images provided by the University of Chicago Library. On the possible editions of the *Cyclopaedia* used by the *encyclopédistes*, see (Passeron, 2006). On Text-PAIR, see https://github.com/ARTFL-Project/text-pair.

[2] See Clovis Gladstone, Russ Horton, and Mark Olsen, "TextPAIR (Pairwise Alignment for Intertextual Relations)", ARTFL Project, University of Chicago, 2008-2021, and, more specifically, (Olsen, Horton and Roe, 2011).

[3] The question of the *Dictionnaire de Trévoux* is one such factor, as it is known that both Chambers and the *encyclopédistes* used it as a source for their own articles—so matches we find between the Chambers and *Encyclopédie* may indeed represent shared borrowings from the Trévoux and not a translation at all. Or, more interestingly, perhaps Chambers translated a Trévoux article from French to English, which a dutiful *encyclopédiste* then translated back to French for the *Encyclopédie*—in this case, which article is the 'source' and which the 'translation'? For more on these particular aspects of dictionary-making, see our previous article (Allen et al., 2010) and a response (Leca-Tsiomis, 2013).

Bibliography

Allen, T. et al. (2010). Plundering philosophers: identifying sources of the Encyclopédie", J *ournal of the Association for History and Computing* **13.1**.

Leca-Tsiomis, M. (2013). The use and abuse of the digital humanities in the history of ideas: How to study the *Encyclopédie*. *History of European Ideas* **39.4**: 467-76.

Lough, J. (1980). The *Encyclopédie* and the Chambers' *Cyclopaedia*. *SVEC* **185**: 221-24

Passeron, I. (2006). Quelle(s) édition(s) de la Cyclopœdia les encyclopédistes ont-ils utilisée(s) ? *Recherches sur Diderot et sur l'Encyclopédie* **40-41**: 287-92.

Olsen, M., Horton, R. and Roe, G. (2011). Something borrowed: Sequence alignment and the identification of similar passages in large text collections. *Digital Studies / Le Champ numérique* **2.1**.

Quintili, P. (1996). D'Alembert 'traduit' Chambers. Les articles de mécanique de la *Cyclopædia* à l'*Encyclopédie*. *Recherches sur Diderot et sur l'Encyclopédie*, **21**:75-90.

Establishing parameters for stylometric authorship attribution of 19th-century Arabic books and periodicals

Romanov, Maxim
maxim.romanov@uni-hamburg.de
Universität Hamburg, Germany

Grallert, Till
t.grallert@fu-berlin.de
Humboldt-Universität zu Berlin, Germany

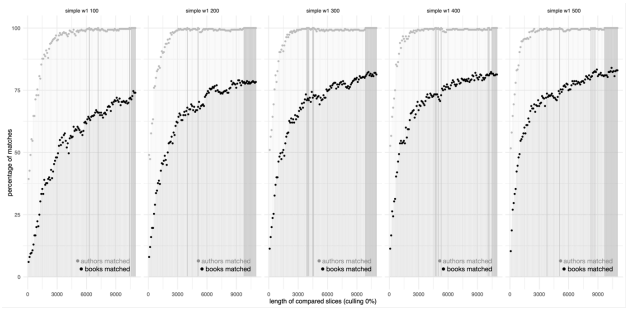
The vast majority of articles in Arabic periodicals from the late Ottoman Eastern Mediterranean (c.1850–1918) carried no explicit authorship information (Grallert 2021, Khayat 2019). Yet, the question of authorship has not received much attention in existing scholarship and is strikingly absent from Ayalon (1995), the standard work in the field. The common implicit hypothesis considers editors-cum-owners listed in mastheads and imprints as the sole authors of all the anonymous texts. This results in the conflation of periodicals with the intellectual output of a single person. Such a synonymous use of, for example, "Muḥammad Kurd ʿAlī" (1876–1953) and the monthly "*al-Muqtabas*" (published in Cairo and Damascus, 1906–1918) can be observed across the board (e.g. Seikaly

1981, Ezzerelli 2017). However, the hypothesis a) remains empirically untested, b) negates the known realities of periodical production and individual biographies, and c) ignores specific contexts of individual periodicals.

Computational stylistics or stylometry is a well-established approach in linguistics and literary studies for authorship attribution and genre detection for major languages of the Global North and has been successfully applied in English and German periodical studies (Benatti and King 2017; Kestemont, Martens, and Ries 2019). "Style", in this context, refers to patterns in the distribution of most frequent linguistic features (most commonly, token or character n-grams). This pattern can be captured statistically and then used to identify authorship of specific texts with high accuracy. This identification works through clustering texts by their similarity according to a variety of distance measures (for example, delta, cosine, euclidean, manhattan, etc.; see, Burrows 2002; Eder 2015; Koppel, Schler, and Argamon 2009). The precision of the approach tends to improve with the length of analyzed samples and Eder (2015) recommends at least 5,000 tokens as a safe threshold for meaningful attribution of prose in English, German, Hungarian, and Polish.

Arabic is a prime example for severely under-resourced languages and scripts of the Global South in the digital realm. Infrastructures of methods, tools, and funding often treat Arabic as an afterthought. Consequently, the rich textual heritage of Arabic-speaking and Islamicate societies is largely absent from debates in Digital Humanities (Miller, Savant, and Romanov 2018). Yet, it is one of the major languages of human cultural production. Arabic script is the second most common after the Latin alphabet and is used for 14 modern languages. Among them, Arabic is the fifth most common language globally with more than 420 million speakers in 26 countries.

Our paper presents the first systematic test of stylometry as implemented in the "stylo" package for R (Eder, Rybicki, and Kestemont 2016) for the analysis of Arabic texts from the long nineteenth century. Extensive parameter testing aimed at empirically identifying optimal sets of parameters for reliable stylometric authorship attribution of Arabic texts. Romanov designed and implemented exhaustive tests of all possible combinations of key parameters on a control corpus of 300 books from 28 authors from the 19th–early 20th centuries (Romanov 2021). For example, MFF: 100-500 in increments of 100; types of MFF: both tokens and characters; lengths of MFF: from 1grams to 4grams in increments of 1; culling unique features: from 0 to 50% in increments of 10; all 14 distance measures available in "stylo"; lengths of samples: from 100 to 12,000 tokens in increments of 100.



We only used consecutive slices in this experiment, since one of the main questions was "what would be the shortest text for which we may still expect reliable results?" The overall design of the test was simple: a new temporary corpus was automatically generated for each combination of parameters where each text was represented by two slices of set length (i.e. we used 600 slices for each test); then we checked how well we could match together slices from the same books and slices by the same authors using Ward's clustering (`ward.D2` in `hclust`). The results were then graphed to allow for a visual exploration of how the precision of matching changes as we gradually increase slices. The graph above shows the best results, which have been achieved with 100-500 *single tokens* as MFF, no culling, and Eder's Simple Delta as the distance measure. In a nutshell, with these parameters, we can expect almost 100% matching with 200-500 MFF and with slices as small as 2,500 tokens. All other combinations of parameters yielded noticeably worse results, which we will discuss in the presentation.

Grallert then tested the reliability of these parameters against a corpus of six periodicals from Baghdad, Beirut, Cairo, and Damascus with some 6 million tokens (Grallert 2020). This corpus differs from the original test corpus in two important ways: First, periodicals represent composites of large numbers of texts of varying length, genre, and authorship. Second, the vast majority of texts carries no explicit or unambiguous authorship information. Yet, the parameters established by Romanov performed equally well for this corpus. Due to the relatively low minimal sample length of 2,500–3,000 tokens, stylometry allows us to double the number of attributed articles in our corpus and to test the bulk of unattributed text against potential contributors and editors. Our results show that the reality of periodical production was more complex than the above-mentioned hypothesis. While Kāẓim al-Dujaylī can be confirmed as one of the editors of *Lughat al-ʿArab* (Baghdad, 1911–14), we can clearly reject the idea of Muḥammad Kurd ʿAlī as the sole editor-cum-author of anonymous texts in *al-Muqtabas*. We also identify distinct shifts in auctorial voices within periodicals that correspond, for instance, to extended absences of editors from the place

of publication. Finally, we show reliable stylistic signals beyond authorship, such as translators and editors, that shed light on the process of periodical production.

In conclusion, our work shows that stylometry can be reliably applied to Arabic texts for the early decades of Arabic print and periodical production and we provide empirically tested and well-documented baseline parameters for similar applications. Future work will have to show to which extent parameters need to be adjusted for earlier periods.

Bibliography

Ayalon, A. (1995). *The Press in the Arab Middle East: A History*. New York: Oxford University Press.

Benatti, F. and King, D. (2017). Hidden Authors and Reading Machines: Investigating 19th-century authorship with 21st-century technologies. University of Victoria, Canada http://www.sharpweb.org/conferences/2017/ (accessed 7 December 2021).

Burrows, J. (2002). 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, **17**(3). Oxford University Press: 267–87 doi: 10/cm2hbk.

Eder, M. (2015). Does size matter? Authorship attribution, small samples, big problem. *Literary and Linguistic Computing*, **30**(2). Pedagogical University of Kraków, Poland and Polish Academy of Sciences, Institute of Polish Language, Kraków, Poland Oxford University Press: 167–82 doi: 10/ggvhx4.

Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R: A Package for Computational Text Analysis. *The R Journal*, **8**(1): 107–21 doi: 10/gghvwd.

Ezzerelli, K. (2017). The publicist and his newspaper in Syria in the era of the Young Turk Revolution, between reformist commitment and political pressures: Muhammad Kurd Ali and al-Muqtabas (1908-17). In Gorman, A. and Monciaud, D. (eds), *The Press in the Middle East and North Africa, 1850-1950: Politics, Social History and Culture*. Edinburgh: Edinburgh University Press, pp. 176–206.

Grallert, T. (2020). Open Arabic Periodical Editions: A framework for bootstrapped digital scholarly editions outside the global north https://openarabicpe.github.io/ (accessed 7 October 2020).

Grallert, T. (2021). Catch Me If You Can! Approaching the Arabic Press of the Late Ottoman Eastern Mediterranean through Digital History. (Ed.) Lässig, S. *Geschichte Und Gesellschaft*, **47**(1): 58–89 doi: 10/gkhrjr.

Kestemont, M., Martens, G. and Ries, T. (2019). A Computational Approach to Authorship Verification of Johann Wolfgang Goethe's Contributions to the Frankfurter gelehrte Anzeigen (1772–73). *Journal of European Periodical Studies*, **4**(1): 115–43 doi: 10/gnq527.

Khayat, N. (2019). What's in a name? Perceptions of authorship and copyright during the Arabic nahda. *Nineteenth-Century Contexts*, **41**(4). Department of Islamic and Middle Eastern Studies, Hebrew University of Jerusalem Routledge: 423–40 doi: 10/gg5zdh.

Koppel, M., Schler, J. and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, **60**(1): 9–26 doi: 10/bnxj7s.

Miller, M. T., Romanov, M. G. and Savant, S. B. (2018). Digitizing the Textual Heritage of the Premodern Islamicate World: Principles and Plans. *International Journal of Middle East Studies*, **50**(1): 103–09 doi: 10/gg865d.

Romanov, M. (2021). A Corpus of Arabic Literature (19-20th centuries) for Stylometric Tests Zenodo doi: 10.5281/zenodo.5772261.

Seikaly, S. (1981). Damascene Intellectual Life in the Opening Years of the 20th Century: Muhammad Kurd ʿAli and Al-Muqtabas. In Buheiry, M. R. (ed), *Intellectual Life In The Arab East, 1890-1939*. Beirut: American University of Beirut, pp. 125–53.

The Modernisa Project: Orthographic Modernization of Spanish Golden Age Dramas with Language Models

De la Rosa, Javier
versae@nb.no
National Library of Norway, Norway

Cuéllar, Álvaro
alvaro.cuellar@uky.edu
College of Arts & Sciences, Hispanic Studies, University of Kentucky, USA

Lehmann, Jörg
joerg.lehmann@romanistik.uni-tuebingen.de
Romanistic Seminar, Eberhard Karls Universität Tübingen, Germany

Introduction

The application of computational analysis to Spanish literature, and to the Golden Age period (16th-17th centuries) in particular, has grown in interest in recent years (De la Rosa and Suárez, 2016; Cerezo Soler and Calvo Tello, 2019; Demattè, 2019; Fiore, 2020; García Reidy, 2019, Vega García-Luengos, 2021). For most of this research (e.g., stylometry, sentiment analysis), a modern and homogenized orthography is usually preferred (Cuéllar and Vega García-Luengos, 2017-2021a-b). In addition, there is a genuine interest in modernization among historians and literature editors, who would benefit greatly from automatic modernization. Unfortunately, we failed to find such systems for Spanish.[1]Most digitization pipelines apply optical character recognition (OCR) to identify the characters of a text as printed, and traditional philologists transcribe texts as faithfully to the original as possible. While new approaches try to improve the existing OCR systems to produce modernized text directly (Cuéllar, 2021a-b), the vast amount of readily available digitized materials in digital libraries and archives cannot be easily re-processed. In this work, we demonstrate how techniques from natural language processing (NLP) can be employed to transform Spanish texts available with historical orthography (circa 1590–1680) into modern normalized Spanish (RAE 2021).

Methodology

The development of the transformer architecture (Vaswani et al., 2017) caused a paradigm shift in NLP. Transformer-based language models excel at many tasks from coherent narrative generation to question answering, and from any sort of classification task to translation (Brown et al., 2020; He et al., 2021, Liu et al., 2020; Xue et al., 2021a). Unfortunately, creating these models requires billions of words, thousands of hours of computation, and many tons of carbon emissions dropped into the atmosphere (Strubell et al., 2019). The bright side is that once a pre-trained language model (PLM) exists, it can be adjusted (fine-tuned) to a specific downstream task with limited data in a fraction of the time and the resources. In this work, we approach orthographic modernization as a translation task and fine-tune existing language models on a parallel corpus of Spanish Golden Age dramas. The majority of PLMs work with vocabularies that might split words into smaller sub-word units called tokens (Devlin et al., 2019). The more frequent a word appears in the pre-training corpus, the higher the probability of keeping the word intact. Since orthographic modernization is a character-based process, we tested both token-free and token-based PLMs. In particular, we fine-tuned the multilingual versions of text-to-text transformers T5 and ByT5 (Xue et al., 2021, 2022) for translation from 17th-century Spanish to modern Spanish and evaluated the results using the BLEU metric (Papineni et al., 2002). In order to avoid misinterpretations of the translation metric caused by the similarity between 17th-century Spanish and Modern Spanish (Post, 2018), we complemented the metric with the average character error rate (CER) and calculated both metrics for the corpus pairs as our baseline.

Corpus Construction

We built a parallel corpus of Spanish Golden Age theater texts with pairs of Golden Age orthography and current orthography. For the old orthography, we used the *Teatro Español del Siglo de Oro*(TESO) corpus (https://quod.lib.umich.edu/t/teso/), because they present the texts "copied exactly as it is written, with all peculiarities captured –accents, abbreviations, etc." (TESO Editorial Policy, online). For the current orthography, we used the *Corpus de Estilometría aplicada al Teatro del Siglo de Oro*(CETSO), a collection of modern editions of the same and many more texts. We chose 44 dramas by the Golden Age dramatists Juan Ruiz de Alarcón, Pedro Calderón de la Barca, Félix Lope de Vega Carpio, and Juan Pérez de Montalbán. All dramas were published in Madrid and Barcelona between 1614 and 1691 for the first time and were written in verses of similar metrical characteristics. Both corpora were aligned line by line to establish a ground truth for the translation between the different historical varieties of Spanish.

Results

After randomizing all 141,023 lines in the corpus, we split it into training (80%), validation (10%) and test (10%) sets stratifying by play. We then fine-tuned T5 and ByT5 base models on sequence lengths of 256 doing a grid search for 3 and 5 epochs, weight decay 0 and 0.01, learning rates of 0.001 and 0.0001, and with and without a "translate" prompt. Table 1 shows the results on the test set of the best model on the validation set for each model type.

	BLEU	CER
Baseline	48.04	8.95%
T5	79.22	4.48%
ByT5	80.66	4.20%

Table 1. Scores for baseline and the best models on the test set.

While both models perform modernization reasonably well, ByT5 seems to be outperforming baseline and T5. We applied our best model to an unseen play (*Castelvines y Monteses*by Lope de Vega, 1647) and analyzed the errors produced. We discovered that the model is capable of solving some difficult corner cases in typographical marks (e.g., adding initial exclamation marks) and some other tricky words (*cómo*vs *como*, *qué*vs *que*) by leveraging contextual information. However, it struggles with proper nouns that normally would go uppercase (e.g., '*Castelvines'*, 'Monteses').We also discovered some strange artifacts in our ground truth corpus regarding archaisms and homogeneity of spelling that might have impacted the learning of the models (e.g., '*efeto'* should appear as '*efecto'* effect, '*agora'* as '*ahora'* now).

Discussion

While the overall error rate of 4.20% can be regarded as satisfying, the results were only evaluated on the basis of dramas written in verse form in 17th-century Spanish. However, there is a broad range of orthographic variation (Mediavilla, 2007) and it may differ from one publishing house or region to another. Thus, the modernization of historical texts that were not produced in the same conditions as our corpus may lead to poorer results. Finally, we found slight differences in punctuation and spelling in our own corpus, even though the aim of these editions was to use modern normalized Spanish. While some of these undesired effects may be addressed by training at the stanza or greater hierarchical level to capture longer range contextual information, it might also imply significantly higher computing resources, training times, and manual revision.

Conclusion

In this work, we have built a parallel corpus of 44 Spanish Golden Age dramas with text in both 17th-century Spanish and Modern Spanish. We have fine-tuned language models on the task of orthographic modernization and show a significant improvement of token-free models over token-based models and baseline. We analyzed closely the errors produced and assessed possible causes and mitigation formulas. We are also releasing our best model hoping to foster research within the Spanish Golden Age period and to establish an alternative to the current cumbersome approach of transcribing Golden Age texts solely by hand.

Availability

A demo of our system can be found at https://huggingface.co/spaces/versae/modernisa

Bibliography

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., et al. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., pp. 1877–901 https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html (accessed 26 April 2022).

Cerezo Soler, J. and Calvo Tello, J. (2019). Autoría y estilo. Una atribución cervantina desde las humanidades digitales. El caso de La conquista de Jerusalén. *Anales Cervantinos*, **51**: 231–50 doi: 10.3989/anacervantinos.2019.011.

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S. and Smith, N. A. (2021). All That's `Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 7282–96 doi: 10.18653/v1/2021.acl-long.565. https://aclanthology.org/2021.acl-long.565 (accessed 26 April 2022).

Cuéllar, Á.(2021a). "Spanish Golden Age Theatre Prints (Spelling Modernization) 1.0". *Transkribus*.

Cuéllar, Á.(2021b). "Spanish Golden Age Theatre Manuscripts (Spelling Modernization) 1.0". *Transkribus*.

Cuéllar, Á and Vega García-Luengos, G. (2017-2021a). C*ETSO. Corpus de Estilometría aplicada al Teatro del Siglo de Oro*, 2017-2021, http://etso.es/cetso/.

Cuéllar, Á and Vega García-Luengos, G. (2017-2021b). *ETSO. Estilometría aplicada al Teatro del Siglo de Oro*. 2017-2021, http://etso.es/.

De la Rosa, J. and Suárez, J. L. (2016). The Life of Lazarillo de Tormes and of His Machine Learning Adversities: Non-traditional authorship attribution techniques in the context of the Lazarillo. *Lemir: Revista de Literatura Española Medieval y Del Renacimiento*(20). Universitat de València: 373–438.

Demattè, C. (2019). Una nueva comedia en colaboración entre ¿Calderón?, Rojas Zorrilla y Montalbán: 'Empezar a ser amigos' a la luz del análisis estilométrico Universidad de Navarra.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings*

of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–86 doi: 10.18653/v1/N19-1423. https://aclanthology.org/N19-1423 (accessed 26 April 2022).

Fiore, A. Questioni di autorialità a proposito di tre commedie seicentesche: Pedro de Urdemalas tra Cervantes, Lope, Montalbán, Diamante e la scuola di Calderón | Artifara. https://www.ojs.unito.it/index.php/artifara/article/view/3970 (accessed 26 April 2022).

García-Reidy, A. (2019). Deconstructing the Authorship of Siempre ayuda la verdad: A Play by Lope de Vega?. *Neophilologus*, **103**(4): 493–510 doi: 10.1007/s11061-019-09607-8.

He, P., Liu, X., Gao, J. and Chen, W. (2021). DeBERTa: Decoding-Enhanced BERT with Disentangled Attention. https://www.microsoft.com/en-us/research/publication/deberta-decoding-enhanced-bert-with-disentangled-attention-2/ (accessed 26 April 2022).

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M. and Zettlemoyer, L. (2020). Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, **8**. Cambridge, MA: MIT Press: 726–42 doi: 10.1162/tacl_a_00343.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–18 doi: 10.3115/1073083.1073135. https://aclanthology.org/P02-1040 (accessed 26 April 2022).

Post, M. (2018). A Call for Clarity in Reporting BLEU Scores. *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, pp. 186–91 doi: 10.18653/v1/W18-6319. https://aclanthology.org/W18-6319 (accessed 26 April 2022).

Reynaert, M., Gompel, M. van, Sloot, K. van der and Bosch, A. van den (2015). PICCL: Philosophical Integrator of Computational and Corpus Libraries: CLARIN Annual Conference 2015. (Ed.) De Smedt, K. *Proceedings of CLARIN Annual Conference 2015*. Wrocław, Poland: CLARIN ERIC: 75–79.

Sebastián Mediavilla, F. (2007). *Puntuación, Humanismo e Imprenta En El Siglo de Oro*. (Publicaciones Académicas 9). Vigo. Pontevedra [Spain]: Academia del Hispanismo.

Strubell, E., Ganesh, A. and McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 3645–50 doi: 10.18653/v1/P19-1355. https://aclanthology.org/P19-1355 (accessed 26 April 2022).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc. https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (accessed 26 April 2022).

Vega García-Luengos, G. (2021). Las comedias de Lope de Vega: confirmaciones de autoría y nuevas atribuciones desde la estilometría (I). *Talía. Revista de estudios teatrales*, **3**: 91–108 doi: 10.5209/tret.74625.

Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A. and Raffel, C. (2022). ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models. *Transactions of the Association for Computational Linguistics*, **10**(0): 291–306.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A. and Raffel, C. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 483–98 doi: 10.18653/v1/2021.naacl-main.41. https://aclanthology.org/2021.naacl-main.41 (accessed 26 April 2022).

Teatro Español del Siglo de Oro (TESO) Editorial Policy https://quod.lib.umich.edu/t/teso/ed-policy.html (accessed 26 April 2022).

Notes

1. Normalization alternatives exist as part of multilingual toolkits that deal with OCR post-correction (e.g., Reynaert, 2015).

Democratizing Poetry Corpora with Averell

De la Rosa, Javier
versae@nb.no
National Library of Norway, Norway

Díaz, Aitor
adiazm@scc.uned.es

UNED, Spain

Pérez, Álvaro
alvaro.perez@linhd.uned.es
UNED, Spain

Ros, Salvador
sros@scc.uned.es
UNED, Spain

González-Blanco, Elena
egonzalezblanco@faculty.ie.edu
IE School of Human Sciences and Technology, Spain

Introduction

Broadly defined, a corpus is a collection of machine-readable texts that are somewhat representative of a particular reality of scholarly interest (McEnery et al. 2006: 5, Xiao 2012: 147). Corpus creation has been part of the research practices of linguists and philologists for decades, and it has recently entered the computer sciences via the mixture field of natural language processing (NLP). Corpora have become a key resource in the development and evaluation of computer systems that deal with language. As these approaches from NLP are being re-discovered, applied, and enriched within the computational humanities, the making of these corpora and their transformation into structured or plain digital texts is of vital importance. Just in the literary domain, there are arguably thousands of corpora available to download or query. In a comprehensive survey, Xiao (2010) describes over a hundred well-known and highly influential corpora in English and other languages. Smaller corpora for understudied or endangered languages have also recently appeared (see Scannell 2007, Ostler 2008, Cox 2011). Notably, only five corpora in these surveys contained poetry and only one of them was annotated with relevant poetic features. As newer poetic corpora with rich annotations are becoming available, we need a proper tool to uniformly access them.

Averell

Among the characteristics that should guide the building of a corpus (McEnery and Wilson, 2001; Gries and Berez, 2015), two are commonly desired: machine readability and availability to researchers. Unfortunately, even when corpora is made fully available in electronic format, it is often the case that scholars struggle to find a proper way to address their research questions using the ready-made corpora (see e.g., Xiao, 2010). In this sense, Averell is a tool that tries to lower the barrier for researchers interested in the study of multilingual poetry corpora. It provides a unified interface to query, manage, download, and merge corpora of poetic nature in multiple languages based on features relevant for poetry scansion and meter analysis. At its core, Averell is a Python library that connects existing annotated corpora in either JSON, XML, or TEI formats, and makes them available into rich CSV and JSON-lines formats that can be later converted into semantic RDF according to the POSTDATA network of ontologies (González-Blanco et al., 2020). Averell exposes a consistent programming application interface to integrate its functionalities into larger software projects, and it is also packaged as a command line tool for its direct use from the terminal.

Granularity

Averell is structured around two key aspects: the catalogue and its granularity. Each corpus defines a granularity level at which its documents can be split. All corpora support splitting by poems and lines (verses), but a line can also be split into words, and then syllables, for which metrical patterns might be provided. In some cases, stanzas, a set of structural and often semantical units within the poem, are also available. Extra information such as the lengths of verses, the amount of lines per stanza, or the type of rhymes is also added when available. This granular annotation allows scholars to merge different corpora and extract sets of poems that meet specific criteria. For example, a corpus of hendecasyllabic safic verses, or poems for a specific period only at the level of the stanza. Instances of the use of Averell to carry out studies in poetry already exist. De la Rosa et al. (2020, 2021) used Averell to create training and validation datasets to fine-tune transformers-based models and to create rule-based systems to a metrical prediction task.

Catalogue

The current catalogue in Averell (Table 1) contains corpora in Czech, English, French, Italian, Portuguese, and Spanish. A total of 12 corpora with 3,847,739 verses are available to download and remix, with different levels of granularity but all of them annotated to a certain extent.

Since corpora have different sizes, formats, and metrical information, we pre-processed each corpus looking for common metadata tags and structures. We then created reusable parsers to extract the relevant information exposed by Averell. The result is a JSON-lines structure capable of capturing the common details of the different corpora. From

this common intermediate format, Averell is able to produce data in formats suitable for analysis such as CSV, Parquet, XML TEI, and even POSTDATA RDF triplets.

Conclusions

In this work, we have introduced the tool Averell for the management and remixing of annotated poetic corporar in a multilingual setting. We have described its structure and showcased a few of its uses in existing scholarly work. We hope to enrich the tool supporting more formats, better interoperability, a larger catalogue, and an easy-to-use web interface.

Table 1. Corpora available and granularity levels for each

Name	Description	Granularity
Disco v2.1 and v3	The Diachronic Spanish Sonnet Corpus (DISCO) Spanish 15th and the 19th centuries sonnets corpus	stanza line
Sonetos siglo de oro	Spanish 16th and the 17th centuries sonnets corpus (Miguel de Cervantes Virtual Library)	stanza line
ADSO 100	Spanish Golden age sonnet corpus	stanza line
Poesía Lírica Castellana del Siglo de Oro	Golden Age Castilian lyric poetry corpus	stanza line word syllable
Gongocorpus	Luis de Gongora poetry corpus	stanza line word syllable
Eighteenth-Century Poetry Archive	English Eighteenth Century poetry corpus	stanza line word
For Better For Verse	University of Virginia poetry corpus	Stanza line
Métrique en Ligne	Université de Caen Normandie (CRISCO) french poetry corpus	stanza line
Biblioteca Italiana	Italian Medioevo to Novecento poetry corpus	stanza
Corpus of Czech Verse	Corpus of Czech poetry of the 19th and of the beginning of the 20th centuries	stanza line word
Stichotheque Portuguese	Stichotheque project portuguese poetry copus	stanza line

Bibliography

Cox, C. (2011). *Corpus Linguistics and Language Documentation: Challenges for Collaboration*. Brill doi: 10.1163/9789401206884_013. https://brill.com/view/book/edcoll/9789401206884/B9789401206884-s013.xml (accessed 28 April 2022).

De la Rosa, J., Pérez, Á., Hernández, L., Ros, S. and González-Blanco, E. (2020). Rantanplan, Fast and Accurate Syllabification and Scansion of Spanish Poetry. *Procesamiento del Lenguaje Natural*, **65**(0): 83–90.

De la Rosa, J., Pérez, Á., Sisto, M. de, Hernández, L., Díaz, A., Ros, S. and González-Blanco, E. (2021). Transformers analyzing poetry: multilingual metrical pattern prediction with transfomer-based language models. *Neural Computing and Applications* doi: 10.1007/s00521-021-06692-2. https://doi.org/10.1007/s00521-021-06692-2 (accessed 28 April 2022).

González-Blanco, E., Ros Muñoz, S., De la Rosa, J., Pérez Pozo, Á., Hernández, L., De Sisto, M., Díaz, A., Khalil, O., Rodríguez, J. L. and Leguina, L. (2020). Towards an Ontology for European Poetry doi: 10.5281/zenodo.4299645. https://zenodo.org/record/4299645 (accessed 28 April 2022).

Gries, S. Th. (2009). What is Corpus Linguistics?. *Language and Linguistics Compass*, **3**(5): 1225–41 doi: 10.1111/j.1749-818X.2009.00149.x.

McEnery, T. and Wilson, A. (2001). *Corpus Linguistics: An Introduction*. Edinburgh University Press.

Ostler, N. (2008). Corpora of less studied languages. *Corpus Linguistics: An International Handbook*, **1**. Walter de Gruyter Berlin: 457–83.

Scannell, K. P. (2007). The Crњbadcn Project: Corpus building for under-resourced languages. *Cahiers Du Cental*, **5**. Citeseer: 1.

Xiao, R. (2010). Corpus creation. In Indurkhya, N. and Damerau, F. (eds), *The Handbook of Natural Language Processing*. 2nd edition. CRC PRESS-TAYLOR & FRANCIS GROUP, pp. 147–65.

Corpus-Based Language Studies: An Advanced Resource Book *Routledge & CRC Press* https://www.routledge.com/Corpus-Based-Language-Studies-An-Advanced-Resource-Book/McEnery-Xiao-Tono/p/book/9780415286237 (accessed 28 April 2022).

Developing the Japanese Visual Media Graph: An Open Knowledge Graph for Researchers Working on Japanese Anime, Manga and Otaku Culture

Roth, Martin
rothm@hdm-stuttgart.de
Stuttgart Media University; Ritsumeikan University

Pfeffer, Magnus
pfeffer@hdm-stuttgart.de
Stuttgart Media University

Kacsuk, Zoltan
kacsuk@hdm-stuttgart.de
Stuttgart Media University

Metadata analytics is a relatively new approach among the many data-driven methodologies engaged with the analysis of culture (e.g. Manovich, 2020; Michel et al., 2011; Moretti, 2013; Rogers, 2013). Using descriptive metadata for research, however, has a well-established tradition in the field of bibliometrics and in particular scientometrics. One of the reasons for the maturity of those fields of research is the long-standing availability of the data itself.

The Japanese Visual Media Graph (JVMG) project, following in the footsteps of the Databased Infrastructure for Global Games Culture Research (diggr) project, is premised on the idea that there exist rich resources on various cultural subfields compiled by online fan or enthusiast communities. By working with these communities towards integrating these descriptive metadata resources into a single knowledge graph for a specific domain – in this case Japanese visual media such as anime, manga, video games and so on – the project aims to open up new avenues of quantitative analysis for researchers in the field, and at the same time provide a template for building similar resources in other areas of inquiry. Although the creation of open knowledge graphs in the digital humanities and the cultural heritage field specifically is becoming increasingly common (see for example Bikakis et al., 2021; Haslhofer et al., 2018), working with data compiled by online fan or enthusiast communities opens up a rich range of new possibilities for research.

The project, which is funded by the German Research Foundation's (Deutsche Forschungsgemeinschaft, DFG) e-Research Technologies program, will be reaching the end of its first project phase in 2022. After three years of work we present the most important aspects of both the knowledge graph that was created (available at https://mediagraph.link/) and the development process that made it possible.

First, we discuss our data sources, the dimensions of the JVMG knowledge graph, and its coverage. We also present our approach to and results of measuring data quality within the project with an emphasis on the accuracy and completeness of the data. This question is especially important in order to, on the one hand, validate the feasibility of using community compiled data for research; and on the other hand, to be able to provide a clear picture for researchers regarding what to expect in relation to the capabilities and limits of the data.

Second, we explain the most important steps and challenges of the data integration process. Working with heterogeneous data sources and ontologies was made possible by transforming all tabular data sources into an RDF linked data format. Matching the data points between the various ingested data sources was one of the significant technical challenges that the project had to resolve. Thus, our experiences with data matching and how much of it was actually automatable is also discussed. All software tools developed for the data ingestion, processing and matching are made openly available online.

Third, the legal harmonization of the JVMG data, which surprised us with its complexity, is another important aspect of the knowledge graph development that we explain in more detail. Licensing and legal concerns are often an afterthought even in scientific projects. In our case, however, since we are working with heterogeneous data sources and an array of corresponding different licensing practices, the question of how to go about harmonizing the licenses of the various data sources became a central problem. This issue had to be solved for us to be able to open up the knowledge graph to researchers around the globe. Our solution, which we introduce along with the challenges that it had to overcome, involved settling on the Creative Commons BY-NC-SA (attribution, non-commercial, share-alike) 4.0 license as the smallest common denominator and requesting individual license agreements from communities whose licenses were not compatible with it.

Fourth, one of the most important ideas underlying our development process was that it had to be directed by the actual needs of researchers working in the field. In order to implement a development process that could receive regular feedback from the domain experts working with the data we adopted and further refined the Tiny Use Case (TUC) methodology introduced by the diggr project (Freybe et al., 2019). This approach builds on ideas from agile software development practices. Each TUC is a small research project that can be tackled within a three to four months long period. Not only do the TUCs serve as examples for

what type of research questions can be pursued with the help of the JVMG knowledge graph, but they also provide valuable lessons in relation to potential problems in data quality, generate new feature requests for the graph frontend – developed based on the Pubby project –, and help identify further types of data that the domain experts would like to be able to work with in the knowledge graph. Last, but not least, each TUC is an opportunity for the team members working on the IT and library and information science side of the project and the humanities and social science researchers to learn from each other and develop a common language and understanding for setting goals and discussing problems. We provide an overview of the TUCs that were conducted in the project and highlight the way they shaped the development of the graph frontend and the knowledge graph itself.

Our hope is that the JVMG project can not only showcase the power of integrated research resources created from data sources compiled by online fan and enthusiast communities, but also provide an array of potential templates (from data integration, legal harmonization and development workflow solutions) for building similar knowledge graphs in other domains of interest.

Bibliography

Bikakis, A., Hyvönen, E., Jean, S., Markhoff, B. and Mosca, A. (eds) (2021). Special Issue on Semantic Web for Cultural Heritage. *Semantic Web*, 12(2).

Freybe, K., Rämisch, F. and Hoffmann, T. (2019). With Small Steps to the Big Picture: A Method and Tool Negotiation Workflow. In Krauwer, S. and Fišer, D. (eds), *Proceedings of the Twin Talks Workshop at DHN 2019* (CEUR Vol-2365). Aachen: CEUR-WS.org, pp. 13-24.

Haslhofer, B., Isaac, A. and Simon, R. (2018). Knowledge Graphs in the Libraries and Digital Humanities Domain. In Sakr, S. and Zomaya, A. (eds), *Encyclopedia of Big Data Technologies*. Cham: Springer, pp. 1-8. doi: 10.1007/978-3-319-63962-8_291-1.

Manovich, L. (2020). *Cultural analytics*. Cambridge, MA: MIT Press.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A. and Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014): 176-182.

Moretti, F. (2013). *Distant reading*. London: Verso.

Rogers, R. (2013). *Digital methods*. Cambridge, MA: MIT press.

Digital Dating and its Discontents: AI, Masculinity and Consent

Roy, Dibyadyuti
dibyadyutir@gmail.com
Indian Institute of Technology Jodhpur ,India,

Dahiya, Lavanya
dahiya.3@iitj.ac.in
Indian Institute of Technology Jodhpur ,India,

Dahiya, Vasundhra
dahiya.2@iitj.ac.in
Indian Institute of Technology Jodhpur ,India,

Abstract

Tinder's Future of Dating Report released in 2021 and purportedly incorporating a survey of 2000 single individuals from India, unequivocally asserted that AI-enabled digital dating had radically altered conversations around consent during the pandemic. Incidentally only a year earlier in 2020, India had been outraged by the *Bois Locker Room i*ncident: wherein an *Instagram* group chat created and disseminated by 27 teenage school students, all male, had circulated nude photos of their female classmates without their consent, along with offensive comments, including mentions of gang rape and other heinous sexual offenses. While igniting heated—albeit short-lived—debates about the ethical and social responsibilities of algorithmically governed platforms these conversations failed to acknowledge how algorithmic cultures and structurations have become embodied in our desiring selves and gendered lives. Therefore, sweeping simplifications about the nature of consent in India emerging from an organization that monetizes desire through an AI-enabled platform environment—where users' agency is operationalized only within the boundaries of that platform —requires an urgent interrogation into both the lack of transparency within such algorithmic systems as well as the concomitant fallacy of stripping algorithmic outcomes from their local specificities and contexts. Particularly in India where (postcolonial) masculinities exist within an irreconcilable field of binaries, between the histories of colonial emasculation and the current hypermasculine machinations of a neoliberal nation (Roy 2021), the algorithmic identities of users while masked under apparent technologized objectivity, are always inflected by complex socio-historical realities (Motihar 2017).

While digital outlets to dating in Global South contexts (like India) should seemingly offer agency to female subjectivities in Gen Z (18–25-year-olds), there is unfortunately little proof that virtual sites of intimacy like Tinder challenge the heteronormative patriarchy and regressive gender roles legitimized through offline social traditions. We chose *Tinder's Future Dating Report* (2021) as an initial point to dive into the algorithmic lives, afterlives, and harms of digital dating platforms for two primary reasons. Firstly, the report contextualises users' engagement with the dating application amidst the Covid-19 pandemic with a particular focus on consent, which the report posits as having become "more commonplace." (Tinder, 2021). Secondly, we operationalize this report as a primer to interrogate the purported premise about the next "decade of [algorithmic] dating" which is posited to be "more honest and authentic." (Tinder, 2021). As gendered subjects of the digital age from the Global South, we found this assertion provocative and a particularly useful entry point into the complex history of gender relations and masculinities in the Indian subcontinent.

In operationalizing the Critical Participatory Inquiry methodological framework (Orlando Fals-Borda 2001) through snowball sampling, based on informed consent, we explore the experiences, fear, and apprehensions of eight Indian respondents in the age group 19-28 regarding digital dating on algorithmically governed platforms. In Indian context, it can be challenging for people to talk about dating experiences. Due to the intimate nature of the research and data, such a sampling method made it easier for respondents to share this data simultaneously allowing us to traverse experiences with greater depth. We analyze the responses received being particularly alive to the socio-economic, class and gendered contexts of our respondents, which thereby inform our discussion into the gendered and intersectional dimensions of algorithmic harms.

Bookended by Global South masculinity studies on one end (Srivastava ; Dasgupta; Kabesh; Roy) and algorithmic accountability scholarship on the other (Acemoglu; Hoffman; Katell et.al; Metcalf et. al), this pilot project explores how AI based digital dating platforms mediate the troubling legacies of hegemonic masculinity in postcolonial spaces, while simultaneously erasing both the gendered anxieties and the human biases from the constituent algorithmic processes. Finally, in a polemic conclusion, we discuss the potential of our project to be an exemplar of critical and decolonial methodologies for examining algorithmic harm and propose paths forward for democratizing conversations around algorithm-driven platforms.

Bibliography

Acemoglu, Daron. 2021. "Harms of AI." Working Paper 29247. Cambridge, MA: National Bureau of Economic Research.

Borda F. , Orlando, Delgado O., and Sandoval R. P. 2009. "Sociólogos o 'sociólogos': La polémica de 1959." Revista Colombiana de Sociología 32(2): 45–60. https://revistas.unal.edu.co/index.php/recs/article/view/12713.

Hoffmann, A. L. (2018, August 25). Data Violence and How Bad Engineering Choices Can Damage Society. Medium. https://medium.com/s/story/data-violence-and-how-bad-engineering-choices-can-damage-society-39e44150e1d4

Kabesh A. T.(2013). "Postcolonial Masculinities: Emotions, Histories and Ethics (The Feminist Imagination - Europe and Beyond)". Ashgate.

Katell, M., Young M., Dailey D., Herman B., Guetler V., Tam A., Binz C., Raz D., and Krafft P M. 2020. "Toward Situated Interventions for Algorithmic Equity: Lessons from the Field," 11.

Metcalf, Jacob, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. "Algorithmic Impact Assessments and Accountability: The Co-Construction of Impacts," ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3442188.3445935

Roy, D. (2020, May 26). Decoding Digital Desires: Postcolonial Masculinity in the Bois Locker Room. The Wire. Retrieved February 6, 2022, from https://thewire.in/women/digital-desires-postcolonial-masculinity-bois-locker-room

Srivastava, S. (2014, December 31). "Sane Sex," the Five-Year Plan Hero and Men on Footpaths and in Gated Communities. On the Cultures of Twentieth-Century Masculinity. Academia.edu. https://www.academia.edu/9967093/_Sane_Sex_the_Five_Year_Plan_Hero_and_Men_on_Footpaths_and_in_Gated_Communities_On_the_Cultures_of_Twentieth_Century_Masculinity

Tinder Newsroom. (2021). The future of dating is fluid. https://www.tinderpressroom.com/futureofdating

The Interpretation of Dreams: A Case Study in Virtual Reality Filmmaking and the Remediation of Psychoanalytic Theory

Sack, Graham Alexander
graham.sack@gmail.com
Washington University in St. Louis, Interdisciplinary Project in the Humanities; Johns Hopkins University, Immersive Storytelling & Emerging Technologies, Film & Media MA Program

This presentation will demo and discuss the production of *The Interpretation of Dreams,* a four-part virtual reality episodic series that immerses users in visually rich and psychologically complex dreamscapes adapted from Freud's original psychoanalytical case studies, including "The Ratman," "Anna O.," "Dora," and "Irma's Injection." The series, released in 2018 during Tribeca Film Festival's Immersive program and subsequently screened at Vancouver International Film Festival and VR Para Llevar, was supported and distributed by Samsung through an experimental grant-making program entitled "VR Pilot Season," the goal of which was to incubate and test the viability of complex, multi-episode, serialized narrative in virtual reality.

The narrative language of two-dimensional cinema co-evolved with the psychoanalytic language of dreams and the unconscious beginning in the early 20th century. Federico Fellini famously called film "a dream we dream with our eyes open," while filmmakers from Méliés to Tarkovsky to Lynch to Nolan have utilized cinema to directly represent dream-states. Dreams inform both the form of these films —their so-called "oneiric" quality, evoking a free-floating, disembodied experience, like drifting through a dream—and their content, which explicitly depicts dreamscapes and the structure of the unconscious.

In many ways, virtual reality as a medium is better adapted to the representation of dream-states and the unconscious than traditional two-dimensional cinema. Most of us do not dream within a frame, after all—we dream immersively. Virtual reality provides a vast new vocabulary for the exploration and visualization of the unconscious, from the construction of surreal landscapes; to the distortion of time, space, perception, and physical law; to user interaction with objects that reveal layers of hidden meaning. Moreover, the language of virtual reality storytelling is still in its infancy, much the way two-dimensional cinema was in the early 20th century.

This project therefore began with the question: "Can the psychoanalytic language of dreams provide guidance and inspiration to immersive filmmakers today, as it did for their predecessors a century ago?"

With this motivating question in mind, the project took as its subject matter the West's most canonized source—Sigmund Freud's *The Interpretation of Dreams* (1900) and the case studies he composed over the successive decade applying his theory. Each episode reimagines one case study as a visually rich, psychologically complex, and emotionally haunting immersive dreamscape.

Episode 1, "The Ratman": A polite but troubled law student arrives at Freud's office complaining of the first recorded symptoms of obsessive-compulsive disorder. As Freud places the patient under hypnosis, the viewer enters a Kafka-esque dreamscape centered around his phobias and sense of criminality.



Episode 2, "Anna O.": The first patient in history to undergo psychoanalysis. Freud described Anna as "the actual founder of the psychoanalytic approach." A 21-year-old paralytic with a rich imagination, Anna expressed her inner world through poetic but melancholy fairy tales inspired by Hans Christian Anderson.



Episode 3, "Dora": Freud's most famous patient, "Dora" was the first test case for the theory of dream interpretation. Arriving at Freud's office suffering from aphonia (the inability to speak), Dora's dreams centered around recurring images of fire, the incineration of her childhood home, and the sinister arrival of "Herr and Frau K."



Episode 4 , "Irma's Injection": In this case, Freud turned his interpretative method back on himself, dissecting the symbolic structure of his own dreams. This episode depicts Freud's own unconscious fears and his lurking sense that the psychoanalytic enterprise may be built on misguided assumptions about the relationship between mind and body.



Each episode was deeply grounded in the source material, but took creative license to render the case in a visually immersive form. One of the central challenges of remediating the psychoanalytic materials arose from the fact that Freud's talking method relied on verbal expression and linguistic association. Virtual reality, even more than traditional cinema, is, however, a visual and sensory medium. It was therefore necessary to find ways to represent the symbolic structure of the cases through visual effects, raising both practical and theoretical questions regarding adaptation, digital remediation, and the narratology of immersive experiences.

The four episodes can be accessed at the following links and either viewed in equirectangular format or side-loaded into a compatible virtual reality headset:

https://www.dropbox.com/s/7catl5l4lx8snrx/dreams_ep1_ratman_mod_fix_mux_360.mp4?dl=0

https://www.dropbox.com/s/9hgsq7rvxdm4x4h/dreams_ep2_anna_mux_360.mp4?dl=0

https://www.dropbox.com/s/irteseie52edlqk/dreams_ep3_dora_mux_360.mp4?dl=0

https://www.dropbox.com/s/f6it65asf5k9syr/dreams_ep4_the_doctor_mux_360.mp4?dl=0

Comparing Symbolism Across Asian Cultural Contexts Using Graph Similarity Measures

Sartini, Bruno
bruno.sartini3@unibo.it
University of Bologna, Italy

Vogelmann, Valentin
valentin.vogelmann@dh.huc.knaw.nl
KNAW Humanities Cluster Amsterdam, The Netherlands

Van Erp, Marieke
marieke.van.erp@dh.huc.knaw.nl
KNAW Humanities Cluster Amsterdam, The Netherlands

Gangemi, Aldo
aldo.gangemi@unibo.it
University of Bologna, Italy

Introduction

Symbols are an essential part of cultures as means to express ideas, values, traditions and as instantiations of belief systems (Kroeber and Kluckhohn, 1952; Brislin, 1976). Unsurprisingly, thus, symbols form the basis of a variety of comparative cultural studies such as evoked concepts in jewellery and ornaments (Zavvāri* and Chitsāziyān, 2021), rituals, mottos, and icons (Manners, 2011), symbolism of trees, dragons, and tree of life (Rival, 2020; Yuan and Sun, 2021; Reno, 1977).

Guided by Martinho (2018), who argues for a shift in cultural studies towards quantitative approaches, and Zepetnek (1999), who adapted comparative literature methodology to identify parallels between cultures, we propose a computational approach that uses symbols for quantified comparative cultural analyses. Leveraging information contained in HyperReal (Sartini et al., 2021), a novel database of symbolism, we define two quantitative measurements of cultural similarity which we apply to its data.

Focussing on a set of cultural contexts from the continent of Asia, and using the defined similarity measures, we address two research questions:

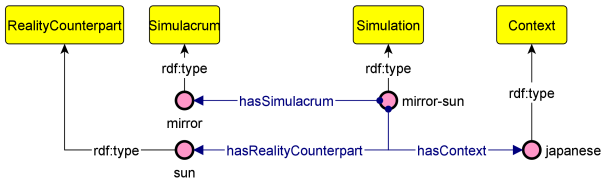
1. Does symbolism provide a useful basis for quantitative cultural comparisons? That is, to what extent does it reproduce expected similarities, and does it have the potential to highlight new, unexpected connections?
2. Do cultures tend to be more similar in terms of their symbols or in terms of the symbolic meanings that their symbols refer to?

For RQ1, we analyse the values of our similarity measures and the clusters of cultures they induce. Additionally, we contrast the similarities of Asian cultural contexts among themselves with two European cultural contexts: Christian and Greco-Roman.

For RQ2, we analyse how similarity values change when applied to either only symbols or only symbolic meanings, as they exist independently in HyperReal.

Linking symbols

Symbolic knowledge has previously been modelled in a semantic web context by (Sartini and Gangemi, 2021) and (Sartini et al., 2021) resulting in the creation of HyperReal, a multi-cultural knowledge graph containing more than 40,000 symbolic meaning relationships (simulations), following the Simulation Ontology schema [1]. In this ontology, symbols (simulacra) are linked to their symbolic meanings (reality counterpart) through an n-ary relationship class Simulation. Simulations are also linked to one or more cultural contexts. Figure 1 shows the example of a mirror (simulacrum), that, in the Japanese context, symbolises the sun (reality counterpart) using HyperReal's structure.



Mirror-sun simulation example

Data selection and extraction

From HyperReal, we selected the 15 unambiguously Asian contexts with the highest number of simulations: Ainu, Buddhist, Cambodian, Chinese, Hindu, Indic, Jain, Japanese, Kalmyk, Mongolian, Philippine, Taoist, Tibetan, Vietnamese, Zoroastrian. This set comprises various types of cultural contexts, such as nationalistic (Chinese) or religious-philosophical (Buddhist), and includes intricate relationships (e.g., Chinese and Taoist). Anticipating that these aspects would emerge from our quantitative analyses themselves, we treat all contexts as equivalent and perform direct comparisons. After the selection, we extracted the subgraphs containing the simulations associated with each context along with the labels for their simulacra and reality counterparts.

Measuring similarity

Semantic approach

Being embodied by linguistic expression allows us to measure symbols' and their symbolic meanings' semantic similarity, for which we use the spaCy [2] and Wiki2Vec (Yamada et al., 2020) Python implementations. [3] We then use the Jaccard similarity metric (Jaccard, 1901) to aggregate sets of semantic similarities for a given pair of cultures. Additionally, we apply weighting according to **symbolic impact** and **symbolic referencing**, where we define **symbolic impact** as the number of symbolic meanings associated with a symbol in a specific context and **symbolic referencing** as the number of times a symbolic meaning is denoted by a symbol in a specific context.

Structural approach

We use graph edit distance (Hagberg et al., 2008) to compute the structural similarity [4] of the extracted cultural contexts' subgraphs. This measurement provides an interface into the similarities of how cultures structurally

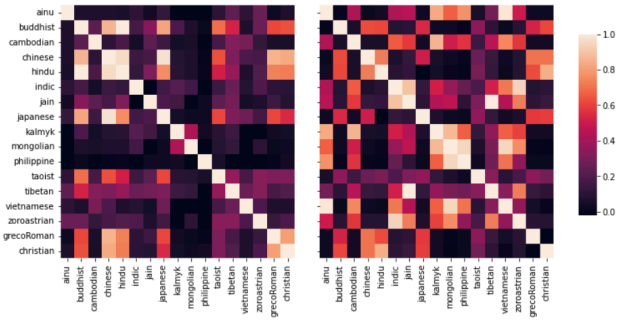
organise symbols, agnostic of the semantics of symbols, and is thus complementary to the semantic approach.

Results

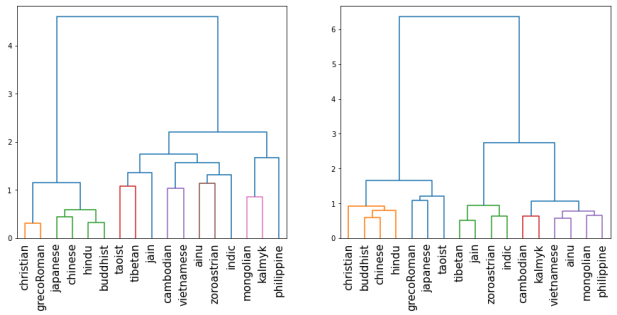
As displayed by Figure 2, our measurements lead to an intricate overall picture of cultural similarities. Whereas, for instance, a larger culture like Buddhist is similar to Chinese, Hindu, Japanese and Taoist; smaller ones like Ainu or Kalmyk are distant, especially semantically, from most other cultures.

Semantic similarity generally seems to be the more conservative, and therefore more often intuitively correct, although counterexamples exist: Jain and Indic, two relatively close cultural contexts, are structurally similar but not semantically so. This underlines the complementary nature of both measurements and is mirrored by the clusters induced from the similarity matrices (Figure 3).

Here, too, groupings of cultures are mostly according to intuition although it is clear that quantitative measures require being supplemented with other sources of information. Then again, as exemplified by the Greco-Roman and Christian cultures, which distinguish themselves from these Asian cultural contexts, connections emerge that are worth further investigation.

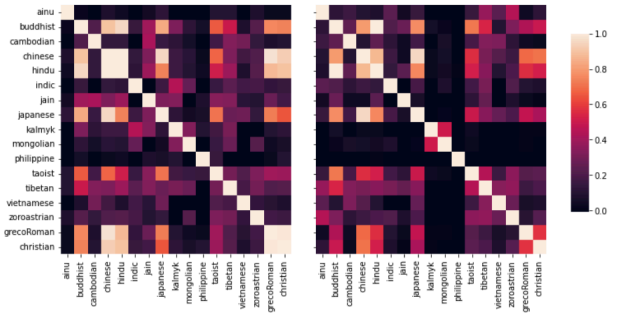


Heatmap of the semantic (left) and structural (right) semantic matrices



Hierarchical clustering induced by semantic (left) and structural (right) similarity matrices. Colours indicate clusters

Regarding RQ2, the investigated cultures, on average, have slightly higher similarities in terms of their symbols than their symbolic meanings. Additionally, Figure 4 shows that some cultures tend to be moderately more similar according to one of the two measurements, such as the Chinese-Jain or the Zoroastrian-Ainu pairs. Thus, cultures seem to not be more similar in terms of either symbols or symbolic meanings, but these have complementary effects to explaining cultural similarity.



Similarity matrix given by the semantics of symbols themselves (left) and symbols‘ meanings (right)

Conclusions

With this work, we initiate quantitative methods for investigations into the similarities of cultures based on symbolism. We provide evidence for their usefulness as a complement to established comparative cultural studies and predict that situating our findings within this field will facilitate new discussions. To this end, future work should also apply the methodology proposed here to larger global sets of cultures to put the similarities within the set of Asian cultures considered here into perspective.

Bibliography

Brislin, R. W. (1976). Comparative research methodology: Cross-cultural studies. *International Journal of Psychology*, **11**(3): 215–29 doi:10.1080/00207597608247359.

Hagberg, A., Swart, P. and Chult, D. (2008). Exploring Network Structure, Dynamics, and Function Using NetworkX.

Jaccard, P. (1901). Etude de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de La Societe Vaudoise Des Sciences Naturelles*, **37**: 547–79 doi:10.5169/seals-266450.

Kroeber, A. L. and Kluckhohn, C. (1952). Culture: a critical review of concepts and definitions. *Papers. Peabody Museum of Archaeology & Ethnology, Harvard University*, **47**(1): viii, 223–viii, 223.

Manners, I. (2011). Symbolism in European integration. *Comparative European Politics*, **9**(3): 243–68 doi:10.1057/cep.2010.11.

Martinho, T. D. (2018). Researching Culture through Big Data: Computational Engineering and the Human and Social Sciences. *Social Sciences*, **7**(12). Multidisciplinary Digital Publishing Institute: 264 doi:10.3390/socsci7120264.

Reno, S. J. (1977). Religious Symbolism: A Plea for a Comparative Approach. *Folklore*, **88**(1). [Folklore Enterprises, Ltd., Taylor & Francis, Ltd.]: 76–85.

Rival, L. (ed). (2020). *The Social Life of Trees: Anthropological Perspectives on Tree Symbolism*. London: Routledge doi:10.4324/9781003136040.

Sartini, B., van Erp, M. and Gangemi, A. (2021). Marriage is a Peach and a Chalice: Modelling Cultural Symbolism on the Semantic Web. *Proceedings of the 11th on Knowledge Capture Conference*. (K-CAP '21). New York, NY, USA: Association for Computing Machinery, pp. 201–08 doi:10.1145/3460210.3493552. https://doi.org/10.1145/3460210.3493552 (accessed 9 December 2021).

Sartini, B. and Gangemi, A. (2021). Towards the unchaining of symbolism from knowledge graphs: how symbolic relationships can link cultures. *Book of Extended Abstracts of the 10th National AIUCD Conference.* Pisa, pp. 576–80.

Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y. and Matsumoto, Y. (2020). Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia. *ArXiv:1812.06280 [Cs]* http://arxiv.org/abs/1812.06280 (accessed 11 December 2021).

Yuan, L. and Sun, Y. (2021). A Comparative Study Between Chinese and Western Dragon Culture in Cross-Cultural Communication. Atlantis Press, pp. 74–77 doi:10.2991/assehr.k.210121.015. https://www.atlantis-press.com/proceedings/sschd-20/125951577 (accessed 24 November 2021).

Zavvāri*, M. and Chitsāziyān, A. (2021). A Comparative Study on the Symbolism in Turkmen and Baluch Ornaments in Iran. *Journal of Iranian Handicrafts*, **3**(2). 30–121 :دانشگاه کاشان.

Zepetnek, S. T. D. (1999). From Comparative Literature Today toward Comparative Cultural Studies. doi:10.7771/1481-4374.1041.

Notes

1. https://w3id.org/simulation/docs
2. https://spacy.io/
3. We measure cosine similarities between vectors of symbols and symbolic meanings generated using the mentioned models.
4. Similarity = 1-graph edit distance

Inner- and Intra-genre Citation Patterns in Film

Schneider, Stefanie
stefanie.schneider@itg.uni-muenchen.de
Ludwig Maximilian University of Munich, Germany

"[A]ny text is the absorption and transformation of another" (Kristeva, 1986) and thus part of a dynamically evolving network of influences. This also applies to film. However, quantitative efforts dissected influence phenomena in film so far only at the individual level (Sreenivasan, 2013; Wasserman et al., 2014; Bioglio and Pensa, 2018). To instead assess inner- and intra-genre citation patterns and thereby challenge hitherto valid genre attributions, we propose the exploitation of three group-dependent network. To this end, genres are considered as fluid, perpetually rearranging concepts (Bordwell, 1989). We leverage user-contributed data from the Internet Movie Database (IMDb). [1] Four types of citations serve as proxies for influence: "edits," "features," "references," and "spoofs"; whereby influence may be an expression of common narratives, settings, or techniques (Lakoff, 1987).

Methods

We regard film citation networks as directed acyclic graphs G=(V,E) consisting of a set of |V|=N films and |

E|=M citations. The directionality of G depends on the order of time: a film x can only cite another film y if y was sufficiently distributed before x. For reasons of simplicity, we assume that G is observed at discrete and equidistant time points t=1,…,T only, so that G(t)=(V,Et) indicates the state of G at time point t. Each state reflects the citations available up until t for films that were produced before t. Over the past decades, numerous metrics have been proposed to determine the influence of such multi-actor collectives in social networks (Silva et al., 2014; Rad and Benyoucef, 2011). We hereinafter focus on three metrics that are interpretable and adaptable, even to humanities-related contexts not directly concerned with citation structures.

Group In-degree and Group Out-degree Centrality

Films are generally considered influential if they are cited by other media; the higher the number of citations received, the greater their perceived importance (Garfield, 1955). In network theory, this is known as *in-degree centrality*. Due to the simplicity of the metric, it is easily extendable to support the analysis of group-dependent citation patterns. As derived from Everett and Borgatti (1999), the *group in-degree centrality* of films associated with genre h can be expressed as the ratio of incoming citations at t from films of other genres, i.e.,

$$\deg_h^-(t) = \frac{\sum_{v \in V_h} \deg_{G^{(t)}}^-(v) - \sum_{v \in V_h} \deg_{G_h^{(t)}}^-(v)}{\sum_{v \in V_h} \deg_{G^{(t)}}^-(v)}$$

Accordingly, the *group out-degree centrality*

$$\deg_h^+(t) = \frac{\sum_{v \in V_h} \deg_{G^{(t)}}^+(v) - \sum_{v \in V_h} \deg_{G_h^{(t)}}^+(v)}{\sum_{v \in V_h} \deg_{G^{(t)}}^+(v)}$$

denotes the ratio of outgoing citations at t to films of genres other than h. Both help to quantify the external and internal willingness of films to establish connections outside their genre, thereby enabling us to trace the genre's historical "evolution" through genre mixing and hybridization.

Group Triangle Participation Ratio

Genres are constructed by a syntax of recurring elements that develop with time (Altman, 1984). The more limited their syntax, the more likely it is that groups with self-referential structures emerge, composed of intra-genre citation triads. We conclude that a film u in such groups not only cites films v and w, but v also cites w. The *group triangle participation ratio*

$$\mathrm{tpr}_h(t) = \frac{\left|\left\{u \;\middle|\; \left\{(u,v),(u,w),(v,w) \in E_h^{(t)}\right\} \neq \varnothing\right\}\right|}{|V_h|}$$

thus is the ratio of citation triads at t of films associated with genre h (cf. Yang and Leskovec, 2012).

Data

To assess genre citation patterns, we retrieve 40,621 feature-length films from the IMDb with a run time of more than 39 minutes that were produced before 2020 and are assigned to at least one of 22 genres. Like Bioglio and Pensa (2018), we observe a high number of films labeled as "drama" or "comedy," with this categorization supported by other labels in 77.938 % and 75.656 % of cases, respectively (Fig. 1A). The IMDb defines eight types of citations that can be entered by users in the "connections" section of the website. We confine ourselves to presumably intentional allusions: of the 126,343 citations, references account for 67.778 % and features for 23.071 %, while spoofs and edits occur much less frequently, at 6.599 % and 2.552 %, respectively. Due to the American-centric bias of the IMDb (Wasserman et al., 2014; Bioglio and Pensa, 2018), we concentrate on films from Europe and North America; English-language films dominate our corpus with 70.978 %. Given that English films are seen more frequently worldwide than non-English ones, the fraction of both incoming and outgoing citations is also disproportionate, possibly underestimating the effect of non-English films and genre influences, like the Italian *giallo*.

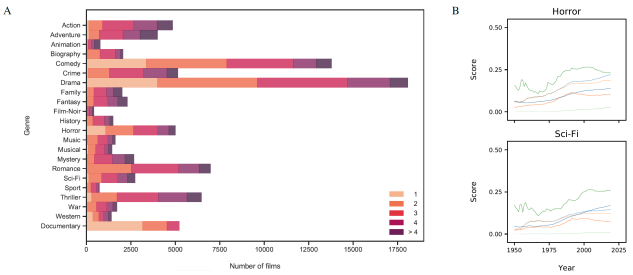


Fig. 1:
(A) Number of films by genre, subdivided by total number of genres stored for films of that genre. (B) Group-dependent metrics over time: group in-degree centrality (blue), out-degree centrality (orange), and triangle participation ratio

(green); with Monte Carlo generated standard deviation bands in lighter colors.

Results

In-degree and out-degree approximately follow power-law distributions, indicating that although most films are cited only few times, a highly influential minority is referenced by dozens (Wasserman et al., 2015; Bioglio and Pensa, 2018). Films that have decisively contributed to the formation or shift of genre identities are particularly important, as they are widely cited not only within but outside their genre (Tab. I).

We determine aforeintroduced metrics for each genre h at time point t. To assess distribution uncertainties, a Monte Carlo simulation with 400 iterations is conducted. At each iteration, genres are randomly redistributed from the underlying empirical frequency distribution. The genre distribution over time remains fixed while the genre distribution at t changes, allowing us to evaluate the approximately true impact of genre-communities. Genres that are predominantly characterized by formulaic narratives or stereotypical figures exhibit large discrepancies between simulated and observed values. For instance, the *group triangle participation ratio* of sci-fi and horror films nearly doubles between 1970 and 2000 with negligible mean simulated ratios (Fig. 1B). This effect is due to few significant genre films that have disrupted conventionalized schemes, triggering new recycling patterns. Standardization through intra-genre repetition seems to be virtually integral to the horror film, e.g., as the lower-than-expected *group in-degree* and *out-degree centrality* confirm. Three citation phases can be verified: gothic horror (*Frankenstein*, 1931), superseded by psychological and occult topoi (*Rosemary's Baby*, 1968) that are increasingly accompanied by slasher films (*The Texas Chain Saw Massacre*, 1974).

Rank	Title	in-degree	$\text{ratio}_{\text{in-degree}_h}$
1	Star Wars: Episode IV (1977)	1190	40.168
2	The Wizard of Oz (1939)	989	35.490
3	Psycho (1960)	624	62.340
4	Jaws (1975)	546	48.718
5	The Godfather (1972)	518	52.703
6	Casablanca (1942)	507	47.732
7	Citizen Kane (1941)	472	51.907
8	Gone with the Wind (1939)	461	52.278
9	The Shining (1980)	455	61.319
10	2001: A Space Odyssey (1968)	443	48.307
11	King Kong (1933)	395	40.253
12	Frankenstein (1931)	395	66.582
13	Taxi Driver (1976)	391	52.941
14	The Terminator (1984)	391	44.501
15	Star Wars: Episode V (1980)	389	48.072
16	E.T. the Extra-Terrestrial (1982)	385	34.545
17	Apocalypse Now (1979)	372	35.215
18	The Exorcist (1973)	369	46.612
19	Raiders of the Lost Ark (1981)	351	45.299
20	Night of the Living Dead (1968)	345	57.681

Tab. 1:
Films by in-degree, with ratio of intra-genre citations in percent.

Bibliography

Altman, R. (1984). A Semantic/Syntactic Approach to Film Genre. *Cinema Journal*, 23.3: 6–18.

Bioglio, L. and Pensa, R.G. (2018). Identification of Key Films and Personalities in the History of Cinema from a Western Perspective. *Applied Network Science*, 3.

Bordwell, D. (1989). *Making Meaning. Inference and Rhetoric in the Interpretation of Cinema*. Cambridge: Harvard University Press.

Everett, M.G. and Borgatti, S.P. (1999). The Centrality of Groups and Classes. *Journal of Mathematical Sociology*, 23.3: 181–201.

Garfield, E. (1995). Citation Indexes for Science. A New Dimension in Documentation Through Association of Ideas. *Science*, 122.3159: 108–11.

Kristeva, J. (1986). *Word, Dialogue and Novel*. In Moi, T. (ed), The Kristeva Reader. New York: Columbia University Press, pp. 34–61.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things. What Categories Reveal About the Mind*. Chicago: University of Chicago Press.

Afrasiabi Rad, A. and Benyoucef, M. (2011). Towards Detecting Influential Users in Social Networks. *International Conference on E-Technologies*, pp. 227–40.

Silva, J.S., de Castro Stoppe, N., Torres, T.T., Ottoboni, L.M.M. and Saraiva, A.M. (2014). Social Network Analysis Metrics and Their Application in Microbiological Network Studies. *Complex Networks V*, pp. 251–60.

Sreenivasan, S. (2013). Quantitative Analysis of the Evolution of Novelty in Cinema Through Crowdsourced Keywords. *Scientific Reports*, 3.

Wasserman, M., Mukherjee, S., Scott, K., Zeng, X.H.T., Radicchi, F. and Amaral, L.A.N. (2014). Correlations Between User Voting Data, Budget, and Box Office for Films in the Internet Movie Database. *Journal of the Association for Information Science and Technology*, 66.4: 858–68.

Wasserman, M., Zeng, X.H.T. and Amaral, L.A.N. (2015). Cross-Evaluation of Metrics to Estimate the Significance of Creative Works. *PNAS*, 112.5: 1281–86.

Yang, J. and Leskovec, J. (2012). Defining and Evaluating Network Communities Based on Ground-Truth. *Proceedings of the IEEE International Conference on Data Mining*, pp. 745–54.

Notes

1. https://www.imdb.com/, accessed 16 March 2022.

Annotating 3D Scholarly Editions

Schreibman, Susan
susan.schreibman@gmail.com
Maastricht University, Netherlands, The

Papadopoulos, Costas
cpapadopoulos84@gmail.com
Maastricht University, Netherlands, The

Annotation has a rich history in the digital humanities: from the by and large manually created text-based annotation found in the Text Encoding Initiative (Cummings 2008) which modelled print-based textual syntax and structures while augmenting the text with commentary; to structured data enabling linked data sources on the WWW in which dynamic annotations are created (Sorbara 2020); from standards such as IIIF which was developed to annotate non-textual electronic files such as audio and video (International Image Interoperability Framework), to automatic image annotation which utilises computer vision to classify and categorise images or parts of images (Abgaz, et al 2021; Wang et al, 2021). The theories, methods, and practices of human-generated annotation have been developed over centuries with its apotheosis, one might argue, in the meticulously researched and richly annotated scholarly editions of the late twentieth century (Shillingsburg 1996). As academics we have been schooled to create and understand the train of scholarship embedded in print-based annotative environments. It is less clear, however, how we are to interpret machine-generated annotation, as well as how the interface and inon-textual annotation play a role in knowledge production and dissemination that is multimodal, interactive, and multisensorial (Apollon, et al 2014; Drucker 2013). These new modalities and environments for annotation raise new issues,such as how to make clear provenance, intention, and context, for both users of these editions, as well as for the scholarly record.

One such new environment that raises such issues is 3D Scholarly Editions. This paper will delve deeper into one aspect of these editions: that of the affordances and challenges in annotating a 3D Scholarly Edition (3DSE) in the form of a virtual world. Previously we have argued that a 3DSE can be likened to a digital edition of an analogue text (such as a novel or a historical document) in which the 3D model is considered the text, with the annotation being part of the apparatus that surrounds it (Schreibman and Papadopoulos 2019; Papadopoulos and Schreibman 2019). However, the metaphor is not a perfect one in that the 3D object is in itself a representation, a domain specific model which simplifies the complexity of the environment being represented. Thus unlike a digital scholarly edition of an analogue text, the “author” of the model (the creator of the 3D representation) can occupy the same role as the “editor”, i.e., the person who annotates and contextualises the model. Even if the modeller and the annotator are different people, the goal of the 3DSE is not to remediate authorial intent (Tanselle 1976) or in this case the intent of the modeller. Rather, the modeller is, we argue, another kind of editor in the text’s (re)construction. To further complicate the actors involved in the 3DSE, given the possibility of more dynamic annotation (e.g., linked data and computer vision), the role of the “editor" can also be assumed by non-human actors.

Therefore, the burden of transparency in indicating agency, intent, and the decisions that the (re)construction is based on is even more complex than in traditional (Digital) Scholarly Editions. Here annotations can take many forms. They can be non-textual, for example, to indicate uncertainty or ambiguity in the (re)construction process when the model created is, for example, of an artefact that no longer exists, or exists in a deteriorated, changed, or incomplete form (such as the reconstruction of a house from antiquity based on its foundations and/or other evidence), or a reconstruction of a church at a particular moment in time when what exists today is an amalgamation of different building phases. In addition to text, annotations can include markup, tags, GIS markers, metadata or paradata that augment understanding, levels of certainty, interpretation, representation or alternative reconstructions, in formats including images, audio, video, 3D, pre-rendered animation, and simulations.

Annotations can also be active or passive. In a 3DSE, an active annotation is triggered by the user based on their interaction within the virtual environment (e.g.,

location, point of view etc.) and can also be personalised depending on the user's interests or research questions. On the other hand, a passive annotation is fixed in space and activated by the user at will. Thus, we argue that within 3D environments, annotations should supply users with the tools to understand the representation (e.g., the model), historically, socially, and/or culturally, as well as the decision-making processes in the creation of the (re)construction.

To explore these issues, this paper will propose a typology of non-mutually exclusive concepts that we consider key to conceptualising annotations and ultimately annotating 3D scholarly environments.

Bibliography

Abgaz, Y., Rocha Souza, R., Methuku, J., Koch, G., & Dorn, A. (2021). A Methodology for Semantic Enrichment of Cultural Heritage Images Using Artificial Intelligence Technologies. *Journal of Imaging*, *7*(8), 121. https://doi.org/10.3390/jimaging7080121

Apollon, D., Bélisle, C., & Régnier, P. (Eds.). (2014). *Digital critical editions*. University of Illinois Press.

Cummings, J. (2008). The text encoding initiative and the study of literature. *A companion to digital literary studies*, 451-76. http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405148641/9781405148641.xml&chunk.id=ss1-6-6, Accessed 26 November 2021.

Drucker, J. (2013). Performative Materiality and Theoretical Approaches to Interface. DHQ: Digital Humanities Quarterly, 7(1). http://www.digitalhumanities.org/dhq/vol/7/1/000143/000143.html

International Image Interoperability Framework, https://iiif.io/

Papadopoulos, C., and Schreibman, S. (2019) "Towards 3D Scholarly Editions: The Battle of Mount Street Bridge." *DHQ: Digital Humanities Quarterly* 13.1. http://www.digitalhumanities.org/dhq/vol/13/1/000415/000415.html

Schreibman, S., and Papadopoulos, C. (2019). "Textuality in 3D: three-dimensional (re) constructions as digital scholarly editions." *International Journal of Digital Humanities* 1(2): 221-233. https://doi.org/10.1007/s42803-019-00024-6

Shillingsburg, P. L. (1996). *Scholarly editing in the computer age: Theory and practice*. University of Michigan Press.

Sorbara, A. (2020). Digital Humanities and Semantic Web: The New Frontiers of Transdisciplinary Knowledge. *Journal of Higher Education Theory & Practice*, *20*(13). http://digitalcommons.www.na-businesspress.com/JHETP/JHETP20-13/16_SorbaraFinal.pdf, Accessed 26 November 2021.

Tanselle, G. T. (1976). The editorial problem of final authorial intention. *Studies in Bibliography*, *29*, 167-211.

Wang, X., Song, N., Liu, X., & Xu, L. (2021). Data modeling and evaluation of deep semantic annotation for cultural heritage images. *Journal of Documentation*, 77(4): 906-925. https://doi.org/10.1108/JD-06-2020-0102

Measuring Space in German Novels

The spatial index (SI) as measurement for narrative space

Schumacher, Mareike Katharina
schumacher@linglit.tu-darmstadt.de
Technical University of Darmstadt, Germany

In this paper, I show how indicators for spatial aspects of narration can be annotated automatically by a German-language machine learning classifier trained on 18th–21st-century novels reaching an overall F1-Score of 0.75 (cf. Schumacher 2021). These indicators can be used as a means of quantification (cf. Bernhart et al. 2018, Kuhn 2018, Schruhl 2018) of narrative space in the form of a spatial index (SI). The formula used for this task sums up the number of annotations per category weighted by explicitness and sets them in relation to the length of the novel. Comparing the SI in a diachronic corpus of 100 German novels, one can detect that indicators for narrative space form a nearly constant part of novels taking up an average of 12% of the content. Compare texts in this way not only opens up a quantitative diachronic perspective on space in novels but also makes it possible to spot outliers that do not fit into this development.

Classifying Space in Novels

Narrative space has been a frequent topic in literary studies (an overview is given by Ryan 2012). Yet it has often been considered less important for narratives than time, because the sequential organisation of events underlying the narrative structure heavily relies on this aspect of human experience (cf. Ryan 2012 §1). The approach presented in this paper takes a first step toward measuring how important the category of space is for narratives. The theoretical foundation used for operationalizing narrative space includes approaches to space from literary studies (e.g. Dennerlein 2009, Piatti

2008, Ryan et al. 2016), Digital Humanities (e.g. Viehauser 2020, Barth and Viehauser 2017, Bodenhammer et al. 2010) and phenomenological studies. The most basic is the distinction between place and space, going back to Descartes (2007;1644). Place is defined as a fixed point in space that can be mapped geographically. Space is understood as a multidimensional continuum whose premiere quality is to be extended. Different conceptions of space were considered when operationalizing narrative space, such as the container space going back to Aristotle (cf. 1995), the geometric euclidean space (Euklid 1971), relational space [1] and space as a socio-cultural phenomenon [2]. In narratives, space can be regarded as a semiotic system, a thematic aspect, or a structural phenomenon of texts (cf. Schumacher forthcoming). Taking operationalizable aspects from all of the above-mentioned approaches to space, I came up with a category system including six categories: place, relations, relational verbs, spatial descriptions, hints on space and topoi.

This category system is implemented in a machine learning process using StanfordNER (cf. Finkel et al. 2005), an implementation of Conditional Random Fields Algorithms (cf. Sutton and McCallum 2007). The machine learning training is organized as an iterative process, including the creation of extensive annotation guidelines (cf. Reiter 2020), manual annotation of training and test data, calculation of an inter-annotator agreement (cf. Artstein 2007, Artstein and Poesio 2008) and an incremental construction on observation of training and test data (cf. e.g. Jannidis et al. 2015, Schumacher and Flüh 2020). The (manually annotated) training corpus consists of 320.000 tokens taken from 80 novels from 18th–21st century. An opening passage of 4.000 tokens of each novel has been integrated. For testing and validating the classifier a leave-one-out-procedure was followed for which 10.000-token passages werde taken from eight novels (two per century) not integrated in the training corpus [3].

A measurement for narrative space

To bring all spatial information into one measurement an index value is calculated, which considers that indicators for representations of space in narratives can be more or less explicit. The most explicit categories are place and relations which are therefore fully counted into the spatial index (SI). Relational verbs are less explicit. For example, when characters in a novel "dance", it is implied that they move, but some other meanings also play an important role. Relational verbs are therefore multiplied by 0.8. The other categories are less and less explicitly referencing space and descend in order: topoi (0.7), spatial descriptions (0.6), hints on space (0.5). The SI is calculated as follows:

$$SI = \frac{\text{places} + \text{relations} + \text{relational verbs} \times 0.8 + \text{topoi} \times 0.7 + \text{spatial descriptions} \times 0.6 + \text{hints on space} \times 0.5}{\text{tokens in the text} \times 100}$$

Space as a categorical constant in novels

Calculating the SI for 100 novels taken from 4 centuries is a good starting point for a diachronic analysis of the representation of space in narratives. As can be seen from figure 1, the representation of space takes up a nearly constant portion of novels from the 18th–21st century with a very slight tendency to rise towards contemporary times.

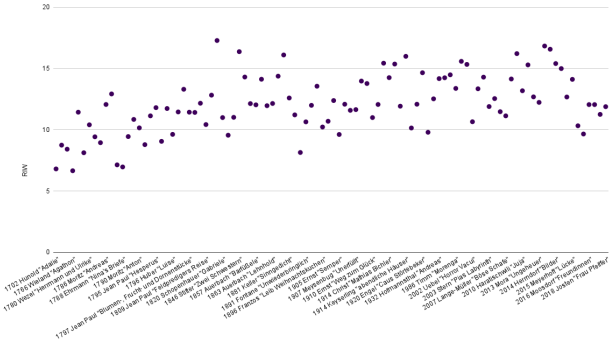


Figure 1:
Spatial indexes in 100 novels from 18th–21st century [4]

What is also striking here is that there is not one single novel without representations of space – in all novels, at least 5% of words carry spatial information. On the other side, there is no novel showing an SI of more than 18. The novel showing the lowest SI (6.66) is *Agathon* by Wieland, the one with the highest (17.29) is *Ahnung und Gegenwart* by Eichendorff. Interestingly in both novels, the protagonists travel, but whereas Wieland's *Agathon* is a Bildungsroman in which the psychological transformation is more important than physical movement, Eichendorff's *Ahnung und Gegenwart* is about a character who severely suffers from war experiences (war being classified as one of the most frequent topoi in novels) and finds his peace in the heterotopic space of a monastery.

Conclusion

Automatic recognition and classification of narrative space in novels can be fruitful for distant reading approaches and help to quantify this aspect of narratives. The calculation of the SI opens up the possibility to compare the representation of space in a diachronic view.

From this perspective, we get first hints on the average representation of narrative space in novels and can spot outliers. Used in this way, the classification and calculation of indicators of narrative space can lead to interesting phenomena in specific texts. For future work it could be interesting to operationalise other basic narrative categories such as time in a similar way. By comparing two or more indexes of such categories would shed more light on the relative importance of space in novels.

Bibliography

Aristoteles (1995): "Physikvorlesung". In: Ders.: *Werke* Band XI. Berlin: Akademie-Verlag.

Artstein, Ron (2017): „Inter-annotator Agreement“. In: Ide, Nancy und Pustejovsky, James: *Handbook of Linguistic Annotation*. Dordrecht: Springer, pages 297–313. DOI: 10.1007/978-94-024-0881-2_11.

Artstein, Ron and Poesio, Massimo (2008): „Inter-Coder Agreement for Computational Linguistics“. In: *Computational Linguistics*, 34(4). DOI:

https://www.mitpressjournals.org/doi/pdfplus/10.1162/coli.07-034-R2.

Barth, Florian und Viehauser, Gabriel (2017): „Digitale Modellierung literarischen Raums“. In: Konferenzabstracts. *DHd2017 Bern. Digitale Nachhaltigkeit*. http://www.dhd2017.ch/wp-content/uploads/2017/03/Abstractband_def3_M%C3%A4rz.pdf [24.2.2020].

Toni Bernhart, Marcus Willand, Sandra Richter, Andrea Albrecht (2018): “Einleitung: Quantitative Ansätze in den Literatur- und Geisteswissenschaften”. In: Bernhart, T., Willand, M., Richter, S. and Albrecht, A. ed. *Quantitative Ansätze in den Literatur- und Geisteswissenschaften: Systematische und historische Perspektiven*. Berlin, Boston: De Gruyter, pp. 1-8. https://doi.org/10.1515/9783110523300-001

Bodenhammer, David J., Corrigan, John und Harris, Trevor M. (2010): *Spatial Humanities. GIS and the Future of Humanities Scholarship*. Indiana: Indiana University Press.

Bourdieu, Pierre (2006; 1989): “Sozialer Raum, symbolischer Raum.” In: Dünne, J.rg und Günzel, Stephan (Hrsg.): *Raumtheorie*. Grundlagentexte aus Philosophie und Kulturwissenschaften. Frankfurt am Main: Suhrkamp, 354-366.

Cassirer, Ernst (2006; 1931): "Mythischer, ästhetischer und theoretischer Raum." In: Dünne, J.rg und Günzel, Stephan (Hrsg.): *Raumtheorie*. Grundlagentexte aus Philosophie und Kulturwissenschaften. Frankfurt am Main: Suhrkamp, 485-500.

Certeau, Michel de (2006; 1980): „Praktiken im Raum.“ In: Dünne, J.rg und

Günzel, Stephan: *Raumtheorie*. Frankfurt am Man: Suhrkamp, 343-353.

Dennerlein, Katrin und Jörg Schönert (2009): *Narratologie Des Raumes*. Berlin; New York: De Gruyter, 2009.

Descartes, René (2007; 1644): *Die Prinzipien Der Philosophie*. Unverändertes eBook der 1. Aufl. von 2007. Hamburg: Meiner. DOI: https://dx.doi.org/10.28937/978-3-7873-2041-7 [29.4.2021].

Euklid (1971): *Elemente*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Finkel, Jenny Rose, Grenager, Trond und Manning, Christopher (2005): „Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling“. In: *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics* (ACL 2005), 363-370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf [5.5.2021].

Foucault, Michel (2008): "Die Ordnung der Dinge" In: Ders. *Hauptwerke*. Frankfurt am Main: Suhrkamp, 7-470.

Husserl, Edmund (2007; 1991): *Ding Und Raum*. Hamburg: Meiner.

Jannidis, Fotis, Krug, Markus, Reger, Isabella, Toepfer, Martin, Weimer, Lukas und Puppe, Frank (2015): „Automatische Erkennung von Figuren in deutschsprachigen Romanen“. In: *Von Daten zu Erkenntnissen*, 1–6. Graz. http://gams.uni-graz.at/o:dhd2015.abstracts-vortraege [5.5.2021].

Kant, Immanuel (2006; 1768): „Von dem ersten Grunde des Unterschiedes der Gegenden im Raum.“ In: Dünne, J.rg und Günzel, Stephan (Hrsg.): *Raumtheorie*. Grundlagentexte aus Philosophie und Kulturwissenschaften. Suhrkamp: Frankfurt/Main, S. 74-76.

Kuhn, Jonas (2018): “Computerlinguistische Textanalyse in der Literaturwissenschaft? Oder: »The Importance of Being Earnest« bei quantitativen Untersuchungen”. In: Bernhart, T., Willand, M., Richter, S. and Albrecht, A. ed. *Quantitative Ansätze in den Literatur- und Geisteswissenschaften: Systematische und historische Perspektiven*. Berlin, Boston: De Gruyter, pp. 11-44. https://doi.org/10.1515/9783110523300-002

Leibniz, Gottfried Wilhelm (2014; 1714): *Monadologie Und Andere Metaphysische Schriften*. Hamburg: Felix Meiner Verlag. DOI: https://dx.doi.org/10.28937/978-3-7873-2117-9 [29.4.2021].

Piatti, Barbara (2008). *Die Geographie Der Literatur*. Göttingen: Wallstein.

Reiter, Nils (2020): „Anleitung zur Erstellung von Annotationsrichtlinien“. In: Reiter, Nils, Pichler, Axel und Kuhn, Jonas (Hrsg.): *Reflektierte algorithmische Textanalyse*. Berlin, Boston: de Gruyter, Seiten 193-202.

Ryan, Marie-Laure, Foote, Kenneth E. und Azaryahu, Maoz (2016): *Narrating Space, spatializing narrative.*

Where narrative theory and geography meet. Columbus: Ohio State University Press.

Schruhl, Friederike (2018): "Quantifizieren in der Interpretationspraxis der Digital Humanities". *Quantitative Ansätze in den Literatur- und Geisteswissenschaften: Systematische und historische Perspektiven*, edited by Toni Bernhart, Marcus Willand, Sandra Richter and Andrea Albrecht, Berlin, Boston: De Gruyter, 2018, pp. 235-268. https://doi.org/10.1515/9783110523300-011

Schumacher, Mareike K. (2021): *Raum-Classifier* (kompatibel mit StanfordNER) (v1.0.0). Zenodo. https://doi.org/10.5281/zenodo.4992662.

Schumacher, Mareike K. (forthcoming): *Orte und Räume im Roman. Ein Beitrag zur digitalen Literaturwissenschaft*. Berlin, Heidelberg: Metzler.

Sutton, Charles, und McCallum, Andrew (2007): *An Introduction to Conditional Random Fields for Relational Learning*. https://people.cs.umass.edu/~mccallum/papers/crf-tutorial.pdf [24.4.2021].

Schumacher, Mareike und Flüh, Marie (2020): "m*w: Figurengender zwischen Stereotypisierung und literarischen und theoretischen Spielräumen. Genderstereotype und -bewertungen in der Literatur des 19. Jahrhunderts." In: Schöch, Christof (2020). *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation*. Konferenzabstracts. DOI: http://doi.org/10.5281/zenodo.3666690 [22.3.2020].

Viehauser, Gabriel (2020): „Zur Erkennung von Raum in narrativen Texten“. In: Reiter, Nils, Pichler, Axel und Kuhn, Jonas: *Reflektierte algorithmische Textanalyse*. Berlin: de Gruyter, 373-390.

Notes

1. Subject-centered space (Leibniz 2014;1714), metaphysical space (Kant 2006;1768), perceptual space (Husserl 2007;1991)
2. Aesthetic space (Cassirer 2006;1931), movement space (de Certeau 2006;1980), action space (Foucault 2008) and milieu (Bourdieu 2006;1989)
3. Training data and test results can be found in the github-repository of the space-classifier https://github.com/M-K-Schumacher/Raum-Classifier
4. Not all novels in the corpus are shown and abbreviations are used. For a full list see the github-repository complementing this work https://github.com/M-K-Schumacher/Forschungsdaten-Orte-und-Raeume-im-Roman/tree/main/Datenbasis_der_Analysen/Raum-Index-Werte

The model of choice.Using pure CRF- and BERT-based classifiers for gender annotation in German fantasy fiction

Schumacher, Mareike Katharina
schumacher@linglit.tu-darmstadt.de
Technical University of Darmstadt

Flüh, Marie
marie.flueh@uni-hamburg.de
University of Hamburg

Lemke, Marc
marc.lemke@uni-rostock.de
University of Rostock

Training a gender classifier for German literature

In several (Digital) Humanities studies, it has been shown that character analysis with a scope on gender can give interesting insides into literary history (cf. Underwood 2019: 111-142, Piper 2018: 133-138). With BookNLP (Bamann 2021) there is a well-performing tool including referential gender inference in the domain of English literary fiction (cf. Underwood 2019: 114). Here, we present a classification tool that is optimized for German fiction and which does not focus on pronouns used for fictional characters but on the ascribed gendered roles (which is referred to as gender identity by Butler 2003). As a starting point, we trained the classifier to annotate the binary (and often stereotyped) gender categories “feminine”, “masculine”, and “neutral”. It is planned to include more categories for gender roles in the future. To reach high accuracy on different literary genres it was trained in an iterative domain adaptation process, which can be roughly split up into three phases (cf. phases 1–3 in table 1).

Phase	Training Set	Ground truth data altogether	average F1-score
1	100.000 tokens taken from 25 19th-century novellas	100.000 tokens (**set 1**)	0.52
2	160.000 tokens taken from 40 novels from 18th–21st century	260.000 tokens (**set 2**)	0.67
3	120.000 tokens taken from 15 plays from 18th–20th century + 7.000 names taken from dramatis personae from 540 plays from 17th–20th century	387.000 tokens (**set 3**)	0.78
4	380.000 tokens taken from training data of phase 3 (without list)	380.000 tokens (**set 4**)	0.79
5	40.000 tokens taken from 10 fantasy novels	40.000 tokens (**set 5**)	0.89
6	420.000 tokens combining ground truth data of phases 4 and 5	420.000 tokens (**set 6**)	0.92

Table 1:
Domain adaptation phases, datasets and performance values (set 1–3: CRF; set 4–6: gBERT)

387.000 tokens, as well as an annotated list consisting of about 7.000 names taken from dramatis personae archived in the GerDraCor-repository (cf. Fischer et al. 2019), have been included in the training corpus. Training data has been manually annotated from scratch, meaning that names and gendered roles have been tagged as either feminine, masculine or neutral. Adding genre-specific training data first leads to an optimization of classification on the specific genre the training data is taken from and second to higher accuracy in the other genres (cf. fig. 1). In the end, our classifier trained with pure CRF algorithms reached 0.86 on novellas, 0.73 on novels and 0.76 on plays. On average the classifier reaches an overall F1 score of 0.78 (Schumacher 2021). The information on gendered roles mentioned in fiction can be combined with other aspects of the analysis of fictional characters such as described features (Schumacher/Flüh 2020), emotions (Schumacher/Flüh 2020, Flüh/Schumacher 2022, Flüh/Horstmann/Schumacher forthcoming) and power structures (Schumacher/Flüh forthcoming).

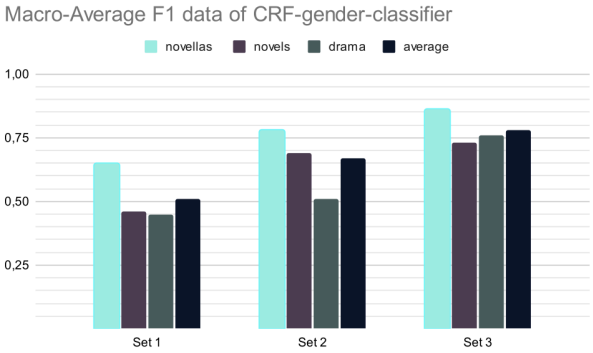


Figure 1:
Training of a generic gender classifier

To adapt recognition and classification to youth fantasy fiction, we tried the implementation of neural networks following a transfer learning approach (cf. phases 4–6 in table 1).

Creating neural net-based Gender Classifiers

We used the software *Neiss TEI Entity Enricher (NTEE),* an implementation of a Transfer Learning (Kamath et al., 2019) approach, to create a neural net-based classifier. Large-scale language models, which are built according to a Bidirectional Encoder Representations from Transformers (BERT) architecture (Devlin et al., 2019) can be fine-tuned using ground truth data for particular NER tasks. In this process, a CRF layer is added to the models (cf. Zöllner et al, 2021: 9–10). For our investigation, we use *gbert-base* (cf. Chan, Schweter, Möller 2021), which is pre-trained with data from the 20th and 21st centuries.

Using sets 4, 5 and 6 (cf. table 1) as ground truth datasets, we compared the performances of classifiers, trained on either generic data, genre-specific data or a combined dataset. Comparing the performance values of the differently designed models shows two things:

1. Using genre-specific data only for fine-tuning the pre-trained gBERT model, in this case, is more efficient than using generic data.
2. Using combined data for fine-tuning the pre-trained gBERT model, in this case, works best.

One can also see a slight difference between the training of the pure CRF-classifier and the fine-tuning of the BERT model (cf. fig. 2). For this implementation, the combination of genre-specific and generic data clearly works best (F1-score of 0.92 tested on fantasy novels).

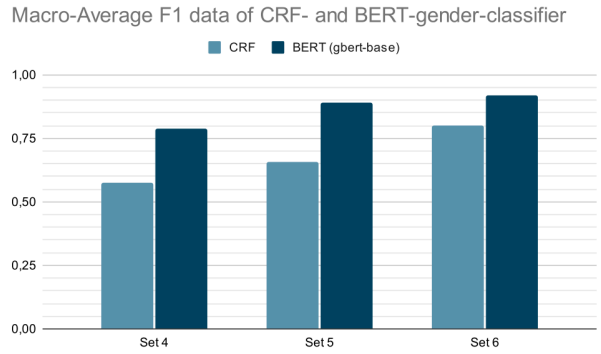


Figure 2:

Average performances of the CRF-based and the gBERT-based classifiers

Bibliography

Beauvoir, Simone de (1992): *Das andere Geschlecht.* Reinbek.

Bourdieu, Pierre (2010): *Die männliche Herrschaft. 1. Aufl., [Nachdr.].* Frankfurt am Main.

Butler, Judith (2003): *Das Unbehagen der Geschlechter. 1. Aufl. [Nachdr.].* Frankfurt am Main.

Chan, Branden, Stefan Schweter, Timo Möller (2020): *German's Next Language Model.* https://arxiv.org/pdf/2010.10906.pdf [Access 8th december 2021].

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "Bert: Pre-training of deep bidirectional transformers for language understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Com-putational Linguistics 1*, p. 4171–4186.

Fischer, Frank, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling und Peer Trilcke (2019): Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. In: Digital Humanities 2019. Utrecht, Zenodo https://doi.org/10.5281/zenodo.4284002.

Flüh, Marie, Jan Horstmann, Mareike Schumacher (forthcoming): Distant Gender Reading Genderaspekte in Fantasy-Jugendromanen von 2008 bis 2020.

Flüh, Marie, & Schumacher, Mareike. (2022, March 7). Jung, wild, emotional? Rollen und Emotionen Jugendlicher in zeitgenössischer Fantasy-Literatur. DHd 2022 Kulturen des digitalen Gedächtnisses. 8. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum" (DHd 2022), Potsdam. https://doi.org/10.5281/zenodo.6327983

Foucault, Michel, Herculine Barbin (2012): *Über Hermaphrodismus: Der Fall Barbin.* Frankfurt am Main.

Kamath, Uday, John Liu, and James Whitaker (2019): *Deep Learning for NLP and Speech Recognition.* Cham: Springer. https://doi.org/10.1007/978-3-030-14596-5.

Manning, Christopher, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky (2014): „The Stanford CoreNLP Natural Language Processing Toolkit". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, p. 55–60, http://www.aclweb.org/anthology/P/P14/P14-5010 [Access 8th december 2021].

Connell, Raewyn (1996): *Gender and power society, the person and sexual politics, Reprint Aufl.* Cambridge.

Connell, Raewyn (2015): *Der gemachte Mann Konstruktion und Krise von Männlichkeiten. Geschlecht & Gesellschaft Band 8, 4. durchgesehene und erweiterte Auflage Aufl.* Wiesbaden.

Piper, Andrew (2018): Enumerations. Chicago: The University of Chicago Press.

Schumacher, Mareike, Marie Flüh (2020): "Figurengender zwischen Stereotypisierung und literarischen und theoretischen Spielräumen: Genderstereotypen und -bewertungen in der Literatur des 19. Jahrhundert". In: Christof Schöch (Hg.): *DHd2020: Digital Humanities zwischen Modellierung und Interpretation*. Konferenzabstracts. o.O. 2020, S. 162–167, https://zenodo.org/record/3666690#.X37-FFlCTus [Access: 26. November 2021].

Schumacher, Mareike (2021): *StanfordNER Gender-Classifier.* DOI: 10.5281/zenodo.5555952.

Schumacher, Mareike, Marie Flüh (forthcoming): "Macht vs. Emotion. Handlungstreibende Muster in Günderrodes Dramen digital, distant und scalable gelesen". In: Roland Borgers, Friederike Middelhoff, Martina Wernli (Ed.): Neue Romantikforschung, Stuttgart: Metzler.

Schumacher, Mareike, Marie Flüh (2021): "Digitale diachrone Korpusanalyse am Beispiel des Projekts „m*w – Gender Stereotype in der Literatur". In: *Digital humanities and gender history*. Jena. DOI: https://doi.org/10.22032/dbt.49173.

Underwood, Ted (2019): Distant Horizons. Chicago: The University of Chicago Press.

Weitin, Thomas (2016): *Volldigitalisiertes XML-Korpus. Der Deutsche Novellenschatz. Hg. von Paul Heyse, Hermann Kurz. 24 Bde. 1871– 1876.* Darmstadt/Konstanz. URL: https://www.deutschestextarchiv.de/novellenschatz/ [Access: 26. November 2021].

Zöllner, Jochen, Konrad Sperfeld, Christoph Wick, and Roger Labahn (2021): *Optimizing small berts trained for german NER*. arXiv. URL: https://arxiv.org/abs/2104.11559 [Access: 8th December 2021].

Systems of Sentencing in Medieval Inquisitorial Records: semantic text modelling as a platform for computational analysis

Shaw, Robert L. J.
robert.shaw@mail.muni.cz
Centre for the Digital Resarch of Religion, Masaryk University, Czech Republic

Outline

This paper shows how the digital capture of texts and quantitative methods can be used to elaborate the weighting of factors that influenced medieval inquisitors in their punishment of religious dissidents. It is a topic which lacks definition in existing research: the extent to which inquisitorial sentencing "systems" existed has escaped qualitative historical approaches. To overcome this, we captured the entirety of a medieval register of inquisition sentences – that of Peter Seila's inquisition in Quercy, Languedoc, 1241–2 – as a series of semantically-rich data statements via a semantic text modelling process developed within the Dissident Networks Project (DISSINET, https://dissinet.cz). From these statements, we then created analytical data projections in order to study the impact of both criminal actions and social connectivity on the penances meted out by Peter through Qualitative Comparative Analysis (QCA) and multiple regression modelling. Overall, our research demonstrates an approach with broader applicability in the digital humanities: one that seeks not only to make texts digitally accessible in the form of richly structured data but also to render every element of both their content and context available to computational study.

Background

In recent decades, there has been significant interest in the way that medieval church authorities in the Latin West approached religious dissidence, above all the way that its representatives construed or even "constructed" heresy (e.g. Moore, 2012; Pegg, 2001). Nevertheless, in-depth modern analyses of the legal processes and sentences that inquisitors – who, from the thirteenth century, took on a key role in prosecuting heresy – employed against dissidents and how these related to the perceived religious and social challenge remain rarer (Given, 1997 and Roach, 2001 remain relatively isolated examples in the context of Languedoc). Crucially the factors that influenced the weight of the punishments the inquisitors handed out have received little systematic attention.

Without understanding these, our knowledge of inquisitorial priorities remains strikingly incomplete. It can be assumed that inquisitors aimed, at least in part, to punish in accordance with the details of the heretical activity that they recorded. But the exact correlations often remain unclear even at the level of a single inquisitor. For instance, were crimes of different types weighted differently? What effect did repetition of crimes have? Inquisitors also captured significant information about the social interactions and relationships of the accused, raising the question of how much they were influenced by these contacts in sentencing. For instance, were dissidents punished differently for knowing famous heretics, or committing certain actions in concert with other suspects? Did kinship ties to other condemned individuals predict graver sentences?

This lack of clarity and definition on these matters results from the methodological and practical research difficulties such questions pose to historians. To answer them, it is necessary both to render sources as structured data in a manner that sufficiently respects the complexity of the cases, and to apply analytical techniques capable of making sense of that same complexity.

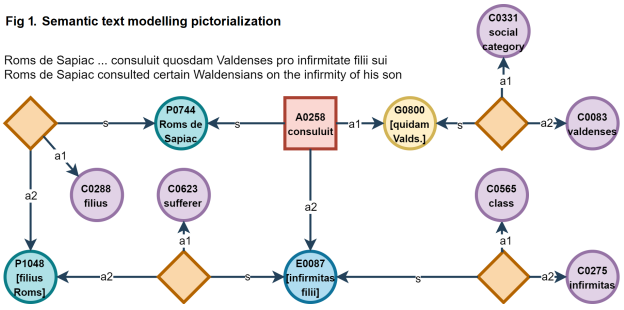
Data and Methods

As a case study, we analysed the register of crimes and sentences drawn up by the Dominican inquisitor Peter Seila in the Quercy region of Languedoc in 1241–2 (Duvernoy, 2001). It is one of the very first extant inquisition registers, recording the sentences of more than 600 supposed religious dissidents accused of involvement in the Cathar and/or Waldensian heresies.

Our data collection from this source was founded on semantic text modelling, a methodology that the DISSINET project has designed for transforming complex textual sources into structured data (Zbíral et al., 2021). Rather than simply isolating features deemed significant for a specific research question, this method seeks to capture almost every detail of textual sources, including their discursive features as well as all evidence of their conditions of production. Each sentence or clause is manually transformed into a "statement" structured as a semantic quadruple (subject, verb, and two object positions). These statements relate "entities" such as Persons, Groups, Concepts, Objects, Locations, Events, and Values (see fig. 1) via an Action. The structured representation of Peter's register in over 10,000 statements (capturing every clause of the source) allowed for a closer feel – both qualitatively, through the process of coding, and quantitatively, through exploratory data analysis – of significant textual patterns before further specifying our analyses.

Fig 1. Semantic text modelling pictorialization

Roms de Sapiac ... consuluit quosdam Valdenses pro infirmitate filii sui
Roms de Sapiac consulted certain Waldensians on the infirmity of his son



With the source digitized in this form, it became possible to transform the data in ways pertinent to different quantitative methods (without any further data entry) in order to study the impact on sentencing of both criminal actions and social connections. By categorizing entities connected to the sentenced individuals within the statements, we created various tables concerning the penances received by these people, their crimes, and the types of people they interacted with and/or were related to by kinship. Using crisp-set Qualitative Comparative Analysis (csQCA) – a technique that uses Boolean minimization to identify sets of conditions that best align with outcomes (Ragin, 2014) – we looked at how the presence and absence of crime types (e.g. resource exchange, ritual, belief, etc.) and the overall sect alignments of the sentenced (i.e. did they interact more with Cathar or Waldensian ministers?) correlate with the essential punishment types (e.g. pilgrimages, crusade service, etc.). To encompass better the repetition of crimes, a broader range of social connectivity conditions, and the complexity of sentencing outcomes, we built a robust multiple linear regression model (Perktold, 2014), using a "Combined Penance Index" (founded on distance of pilgrimages / duration of other sanctions) as the dependent variable. Through this, we sought to understand 1) the linear dependence between the proportions of criminal acts of different types and punishment and 2) the positive influence of the social context of crimes on sentencing (i.e. a "guilt by association" hypothesis).

Findings

The combined results show that Peter Seila was relatively systematic throughout the trials in his weighting of the different types of crimes and interactions with dissident ministers when sentencing: for instance, ritual crimes and Cathar interactions appear particularly associated with severe outcomes. We found no evidence, however, that Peter was influenced by accomplicity or kinship among the sentenced followers. More broadly, our research demonstrates how a text captured in digital form via a semantic text modelling process can be quantitatively analysed through multiple approaches, and with a precision concerning content and context that will satisfy researchers from a qualitative background.

Bibliography

Duvernoy, J. (2001). *L'inquisition en Quercy: le registre des pénitences de Pierre Cellan, 1241–1242*. Castelnaud la Chapelle: L'Hydre Éditions.

Given, J. B. (1997). *Inquisition and Medieval Society: Power, Discipline, and Resistance in Languedoc*. Ithaca & London: Cornell University Press.

Moore, R. I. (2012). *The War on Heresy*. Cambridge, MA: The Belknap Press of Harvard University Press.

Pegg, M. G. (2001). *The Corruption of Angels: The Great Inquisition of 1245-1246*. Princeton: Princeton University Press.

Perktold, J. (2014). REF/ENH RLM and robust scale for almost perfect prediction. *Statsmodels Github*. https://github.com/statsmodels/statsmodels/pull/1341(last modified 2 April 2021; accessed 2 October 2021)

Ragin, C. C. (2014). *The comparative method: moving beyond qualitative and quantitative strategies*. 2nd ed. Los Angeles: University of California Press.

Roach, A. P. (2001). Penance and the making of the inquisition in Languedoc. *Journal of Ecclesiastical History,* **52**: 409–433.

Zbíral, D., Shaw, R. L. J., Hampejs, T., and Mertel. A. (2021). Model the source first! Towards source modelling and source criticism 2.0. *Zenodo*, https://doi.org/10.5281/zenodo.5218926 (accessed 21 April 2022)

Poetry as Error. A 'Tool Misuse' Experiment on the Processing of German Language Poetry

Sluyter-Gäthje, Henny
sluytergaeth@uni-potsdam.de
University of Potsdam, Germany

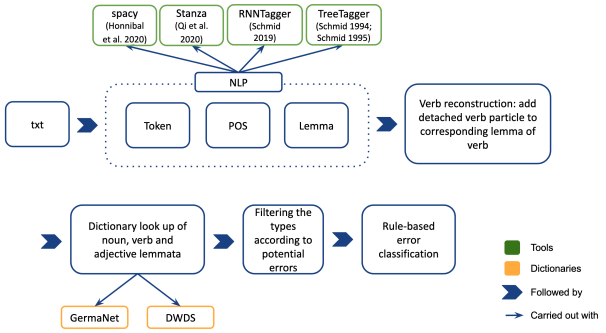
Trilcke, Peer
trilcke@uni-potsdam.de
University of Potsdam, Germany

1. Research question

In Computational Literary Studies texts are typically pre-processed with Natural Language Processing (NLP) tools. However, due to historical and/or aesthetic characteristics, literary texts sometimes deviate notably from the data the tools are trained on. Due to this difference in domain, the performance of the tools drops (Scheible et al., 2011; Rayson et al., 2007; Herrmann, 2018; Bamman, 2020). Instead of considering this to be a problem, the 'erroneousness' of the tools could provide a computational understanding of the 'deviance of literary texts'; produced errors might reveal something about the characteristics of literature.

In the following, we report on a *Tool Misuse* experiment on German lyric poetry – a genre that is usually associated with a high degree of deviance (Müller-Zettelmann, 2000: 100; Zymner, 2019: 29–30) – in which we develop a pipeline that provokes tokenization, lemmatization and POS tagging 'errors' of NLP tools and typologises these 'errors' in a rule-based way.

2. Operationalization



Pipeline for error typing of the corpora.

Since gold standard annotations are not available for our scenario, we base our evaluation on the assumption that correctly produced lemmas can be found in dictionaries of German language. Based on the TextGridRepository, we build a canon-based corpus of 'prototypical' German-language poetry comprising 5,144 poems. For comparison, we use a prose corpus of 100 German-language novels from the 19th century, compiled from the TextGridRepository and Project Gutenberg. As dictionaries we use ‘GermaNet‘ (Hamp & Feldweg, 1997; Henrich & Hinrichs, 2010) and the ‘Digitales Wörterbuch der deutschen Sprache‘ (Klein & Geyken, 2010). To ensure that the resulting errors are not tagger specific, we use several NLP tools for tokenization, lemmatization and POS tagging of the corpora (fig. 01) and consider all content word types as potential errors that are lemmatized by at least two tools and for which none of the produced lemmas are found in the dictionaries (fig. 02, column "pFail").

Our 'error pipeline' prefers recall over precision, thus it produces only circumstantial evidence of potential errors. A larger number of false positives is to be expected, because we process out-of-vocabulary words of the dictionaries.

	Poetry	Poetry	Prose	Prose
	all	pFail	All	pFail
Types	70,422	24,244	263,042	115,785

Number of word types (NOUN, VERB, ADJECTIVE) for the entire corpus ("all") and for the sets with potential errors ("pFail").

3. Analysis

Based on manual inspections of the pFail set, we postulate 13 error types described in figure 03. For each type we formulate a rule [1] which is then applied to the pFail set following the order of the error types listed below. Multiple typings are not possible.

Name	Description
CONTRACT	The pronoun "es" is contracted to the preceding word in the form of "'s".
ELISION_APO	Vowels within a token are replaced by apostrophes.
PUNC	A punctuation mark is pos-tagged as NOUN, VERB, ADJECTIVE or a punctuation mark is tokenised as part of a word.
SHORT	Single letters or digits are pos-tagged as NOUN, VERB, ADJECTIVE.
COMP_DASH	Several NOUNs are joined together with a hyphen to form a word that is not in the dictionary.
COMP	Several NOUNs are joined together to form a word that is not in the dictionary.
PART_ADJECTIVE	VERB is derived to a participial ADJECTIVE that is not in the dictionary.
ELISION_SIMPLE	Vowels are deleted in the penultimate or last syllable of a word without being marked by an apostrophe.
ORTH_UPPER	ADJECTIVE or VERB is capitalised at the beginning of a verse.
ORTH_SZ	A word uses the historical spelling with "ß".
PREFIXED	A prefix is used to derive a word that is not in the dictionary.
EPITHESIS	An "e" is added to NOUN.
ELISION_END	The coda in NOUN is deleted.

Description of error types.

4. Results

	Poetry	Prose
PUNC	0.454	0.576
SHORT	0.223	0.124
ORTH_SZ	0.120	0.133
ORTH_UPPER	0.627	0.020
ELISION_APO	2.083	0.314
ELISION_SIMPLE	2.574	0.978
ELISION_END	1.081	0.246
EPITHESIS	0.285	0.128
CONTRACT	0.639	0.271
COMP_DASH	1.926	4.957
COMP	37.783	46.432
PART_ADJECTIVE	3.234	2.060
PREFIXED	2.302	3.643

Relative frequency for the types of potential errors for the two pFail sets.

53.33 % of the word types in the pFail set for poetry and 59.88 % of the word types in the pFail set for prose are identified. PUNC and SHORT are predominantly sub-word level characters, mostly noise which appears to a comparable extent in poetry and prose. ORTH_SZ reflects the effect of Historical Orthography which a normalisation step could remedy.

The ten remaining types can be combined into three groups:

- COMP_DASH, COMP, PART_ ADJECTIVE, PREFIXED gather *Creative Lexis*, i.e. word formation mechanisms (composition, derivation); these are often out-of-vocabulary words and therefore pipeline errors, not tool errors. In poetry, 45.25 % of the "pFail" set can be assigned to this group, in prose 57.09 %.
- As expected, the pipeline produces a higher error rate for poetry (0.62 %) than for prose (0.02 %) for ORTH_UPPER, which identifies a characteristic of *Lyric Typography* (capitalizing first letters in lines).
- The error rate of *Prosodic Deformation* consisting of ELISION_APO, ELISION_SIMPLE, ELISION_END, EPITHESIS and CONTRACT is also higher for poetry than for prose (6.62 % compared to 1.93 %). We assume that the deformations are due to the addition or deletion of vowels for metric reasons.

5. Outlook

Our pipeline identifies *Prosodic Deformation*, *Lyric Typography* and *Creative Lexis* as typical sources of error when processing poetry with NLP tools. However, our pipeline needs to be optimized: too many potential errors are, as in the case of *Creative Lexis*, in fact not tool errors but pipeline errors. Additionally, our rule-based typology is only able to describe 53.33 % of the pFail set. This reveals two areas for follow-up research: the pipeline could be improved on to decrease the number of pipeline errors and the rule-based typologisation procedure could be optimized against our baseline.

Bibliography

Bamman, D. (2020). LitBank: Born-Literary Natural Language Processing. [Preprint]. https://people.ischool.berkeley.edu/~dbamman/pubs/pdf/Bamman_DH_Debates_CompHum.pdf [Last accessed November 16, 2021].

Braam, H. (2019). *Die berühmtesten deutschen Gedichte. Auf der Grundlage von 300 Gedichtsammlungen*. Stuttgart: 2. Aufl., Kröner.

Hamp, B. and Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid, Spain, pp. 9–15. https://aclanthology.org/W97-0802.pdf [Last accessed November 16, 2021].

Henrich, V. and Hinrichs, E. (2010). GernEdiT - The GermaNet Editing Tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, pp. 2228-35. http://www.lrec-conf.org/proceedings/lrec2010/pdf/264_Paper.pdf [Last accessed November 16, 2021].

Herrmann, J. B. (2018). Praktische Tagger-Kritik. Zur Evaluation des PoS-Tagging des Deutschen Textarchivs. In *DHd2018: Kritik der digitalen Vernunft. Book of Abstracts.* Cologne, Germany, pp. 287-90. https://zenodo.org/record/3684897#.YO_x1W5CTOQ [Last accessed November 16, 2021].

Honnibal, M. et al. (2020). spaCy: Industrial-strength Natural Language Processing in Python. Zenodo. https://doi.org/10.5281/zenodo.1212303 [Last accessed November 16, 2021].

Klein, W. and Geyken, A. (2010). Das 'Digitale Wörterbuch der Deutschen Sprache DWDS', in: Lexicographica 26: 79–96.

Müller-Zettelmann, E. (2000). *Lyrik und Metalyrik. Theorie einer Gattung und ihrer Selbstbespiegelung anhand von Beispielen aus der englisch- und deutschsprachigen Dichtkunst*. Heidelberg, Germany, Winter.

Qi, P. et al. (2018). Universal dependency parsing from scratch, in: *Proceedings of the CoNLL 2018 Shared*

Task: Multilingual Parsing from Raw Text to Universal Dependencies, Brussels, Belgium, pp. 160-70.

Rayson, P. et al. (2007). Tagging the bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In *Proceedings of Corpus Linguistics (CL2007)*. https://eprints.lancs.ac.uk/id/eprint/13011/1/192_Paper.pdf [Last accessed November 16, 2021].

Schmid, H. (1994). Probabilistic part-of speech Tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK, pp. 154-62.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*, Dublin, Ireland, 13-25. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf [Last accessed November 16, 2021].

Schmid, H. (2019). Deep learning-based morphological taggers and lemmatizers for annotating historical texts. *In Proceedings of the 3rd international conference on digital access to textual cultural heritage*, Brussels, Belgium, 133-37.

Scheible, S. et al. (2011). A gold standard corpus of Early Modern German. In: *Proceedings of the 5th Linguistic Annotation Workshop*, pp. 124–28. https://dl.acm.org/doi/abs/10.5555/2018966.2018981 [Last accessed November 16, 2021].

Zymner, R. (2019). Begriffe der Lyrikologie. In: Hildebrandt, Claudia et al. (eds.) Lyrisches Ich, Textsubjekt, Sprecher? (= Grundfragen der Lyrikologie, Bd. 1). Berlin, Germany: De Gruyter 25–50.

Notes

1. For the rules see: https://gitup.uni-potsdam.de/sluytergaeth/poetry_as_error

Digital Humanities and Replication. Ingredients for a Love Story – Experiences from the '(Re)counting the Uncounted' Project

Stapel, Rombert
rombert.stapel@iisg.nl
International Institute of Social History, Netherlands, The

Introduction

Over the past years much has been written about the replication crisis in science, with humanities research slowly, and sometimes hesitantly, catching up (Peels and Bouter, 2018a; Rijcke, de and Penders, 2018; Peels and Bouter, 2018b). Strongly related to this development, is the strive for open science (UNESCO, 2021). By facilitating and compelling researchers (through research grant stipulations) to not only publish their results, but also to provide access to their data and methodologies, a major prerequisite for replication is met. [1] The FAIR principles to data management (Findable, Accessible, Interoperable, Reusable) are central to this approach (Wilkinson et al., 2016).

Current literature on the relationship between digital humanities and the replication of research to a large extent focuses on the challenges of replicating digital, or more broadly speaking, quantitative-based humanities research (Tucker, 2017; Flis, 2018). In computational linguistics discussions have taken off, not least in response to a recent thought-provoking article by Nan Z. Da (Da, 2019; Algee-Hewitt et al., 2019; Arnold and Buell, 2019; Antoniak et al., 2020).

Rather than presenting another gloomy narrative on the replicability of digital humanities research, in this paper we want to address the opportunities that digital humanities methodologies offer in mediating the 'replication crisis.' Although digital humanities are sometimes narrowly defined, barely interacting with quantitative historical research for instance, we will use it as a broad umbrella term for humanities research that systematically makes use of and analyses digital resources. Such research almost always includes some form of empirical analysis, thus lending itself to replication as well (Peels, 2019).

Background of the project

Such an approach is central to the *(Re)counting the Uncounted* project, the first humanities project funded through the Dutch Scientific Council's Replication Studies Program – unique in the world. [2] In this project, four seminal studies which have estimated the medieval and early-modern population in the Netherlands and Belgium are formally replicated (Faber et al., 1965; Blockmans et al., 1980; Klep, 1991; Paping, 2014). This replication is performed by using the same underlying data, many hundreds of premodern censuses, and by applying the same methodologies: usually multiplying the units of the actual censuses – typically hearths, houses, chimneys, families, communicants, able-bodied men, and sometimes individuals – with predetermined coefficients, while accounting for

those who are excluded from the census (for reasons of fiscal exemption for instance) (for an introduction to the challenges with these types of sources: Arnould, 1976).

The *(Re)counting the Uncounted* project aims to test the consistency of the methodologies applied in the four studies by using digital humanities methods as a feedback loop. For this purpose, we have a twofold approach. First, we digitise and contextualise the **unaggregated** – i.e. on village-level – premodern censuses. These statistics are then linked to specially prepared Historical GIS maps of locality boundaries in the Netherlands, Belgium, Luxembourg, and surrounding areas (Stapel, 2020). Second, there is the problem of the diverging nature of the censuses. These censuses have been created by a multitude of actors, in a multitude of territories, for a multitude of purposes, counting a full range of units, and across nearly five centuries. To exacerbate the problem of the comparability of their results, modern scholarship has never been able to reach consensus on which coefficients should be used in what circumstances (e.g.: Blockmans et al., 1980: 42–43; Stabel, 1997: 19 ff.; Hélin, 1963: 41 ff.; Brouwer, de, 1963; Cloet, 1966; Woude, van der, 1972: 77–91).

By carefully building not only a database of digitised censuses, but also a full record of contextual information on each census (what is counted, who is counted, by whom is the census made, for what purpose, etc.), it becomes possible to analyse this contextual information, for instance to create more consistent coefficients. Moreover, it will become possible, through the advancements of GIS techniques, to contextualise the (socio)geographical context of a locality mentioned in one of these medieval or early modern censuses (e.g.: Stapel, 2017: 182).

Facilitating replicability through database design

Thus far, we have focused on the set up of this humanities replication study. The set up however has much wider implications for how, in this case, historical statistical databases should be constructed in our opinion. Here too, digital humanities techniques, although not uniquely developed for or within a humanities context, play an essential role.

A traditional database of historical statistics – again we use this type of database as an example, but it can be applied much broader in humanities research as well – involves a database of, typically, rows and columns – mimicking the printed table well-known in scholarly literature for centuries. Rather than putting the rows and columns at the forefront, and defining them, we aim to put the data observation central. Every data observation is linked to contextual information, which may also include specific information usually stored in table footnotes, and will be geographically defined in GIS. We will use Linked Open Data (LOD) to facilitate this database structure.

One may argue, with reason, that this approach is not very new. Yet, the approach is still very uncommon in quantitative humanities research and rarely applied in full. Building from our experiences in the *(Re)counting the Uncounted* project, we will also show that the amount of time needed to invest in such a data model should not necessarily be an obstacle, nor should lack of access to user-friendly methods to set this data model up. In its very core, the *(Re)counting the Uncounted* project, while promoting all four FAIR principles, aims above all to improve the interoperability of through-and-through messy historical data in this way.

Moreover, and this is also a vital element, in relation to the replication aspect, the contextual and geographical information stored with every data observation can be further distributed to users of the data, creating a new level of transparency. After all, replication does not end with the publication of new results, but involves an ongoing conversation (Peels, 2019). Facilitating the replicability of replicated research is essential. We will exploit the possibilities of LOD to create a crumb trail from an observation in a source (a table in literature, an (image of a) archival document, etc.), via a range of carefully defined interpretations of that observation (either by existing scholars or by future users), to a scientific product: in our case population estimates based on very distinct types of sources.

Finally, in order to facilitate source critical attitudes of the users, we aim to grant access to our (open) data through dynamic questionnaires. These have the purpose of bringing any user up to speed with the specific challenges of our source material, forcing them to think about how these challenges affect their research question. Downloading the aggregated end results in a CSV without ever considering how the data came into existence – a common research practice, at least in quantitative history – is actively discouraged in this way.

Acknowledgments

Co-author to the paper and LOD model is Ivo Zandhuis (Fellow at the International Institute of Social History and independent researcher and consultant at ivozandhuis.nl). This publication is part of the project ‘(Re)counting the Uncounted’ (with project number 401.19.038 of the research programme Replication Studies which is (partly) financed by the Dutch Research Council (NWO).

Bibliography

Algee-Hewitt, M. A., Bode, K., Brouillette, S., Finn, E., Klein, L., Long, H., Piper, A., Underwood, T., Da, N. Z. and Fish, S. (2019). Computational Literary Studies: A Critical Inquiry Online Forum *Critical Inquiry* https://critinq.wordpress.com/2019/03/31/computational-literary-studies-a-critical-inquiry-online-forum/.

Antoniak, M., Jannidis, F., Mimno, D., Schöch, C. and Dalen-Oskam, K. van (2020). Replication and Computational Literary Studies. *DH2020*. Ottawa: ADHO doi:http://dx.doi.org/10.17613/ekd2-ew51. https://hcommons.org/deposits/item/hc:30439/ (accessed 10 December 2021).

Arnold, T. and Buell, R. (2019). More Responses to 'The Computational Case against Computational Literary Studies' *Critical Inquiry* https://critinq.wordpress.com/2019/04/12/more-responses-to-the-computational-case-against-computational-literary-studies/.

Arnould, M.-A. (1976). *Les Relevés de Feux*. (Typologie Des Sources Du Moyen Âge Occidental 18). Turnhout: Brepols.

Blockmans, W. P., Pieters, G., Prevenier, W. and Van Schaïk, R. W. M. (1980). Tussen crisis en welvaart: sociale veranderingen 1300-1500. In Blok, D. P. (ed), *Algemene Geschiedenis Der Nederlanden*, vol. 4. Haarlem: Fibula-Van Dishoeck, pp. 42–86.

Brouwer, J. A. K. de (1963). Het belang van de kommunikantencijfers en de verhouding ervan tot de bevolking. *Handelingen van de Koninklijke Zuidnederlandse Maatschappij Voor Taal- En Letterkunde En Geschiedenis*, **17**: 67–80.

Cloet, M. (1966). De leeftijdsgrens tussen communicanten en niet-communicanten in de XVIIde en de XVIIlde eeuw. *De Leiegouw*, **8**(2): 451–71.

Da, N. Z. (2019). The Computational Case against Computational Literary Studies. *Critical Inquiry*, **45**(3): 601–39 doi:10.1086/702594.

Faber, J. A., Roessingh, H. K., Slicher van Bath, B. H., Van der Woude, A. M. and Xanten, H. J. van (1965). Population changes and economic developments in the Netherlands: a historical survey. *A.A.G. Bijdragen*, vol. 12. Wageningen: Afdeling Agrarische Geschiedenis, Landbouwhogeschool, pp. 47–113.

Flis, I. (2018). Digital humanities as the historian's Trojan horse: Response to commentary in the special section on digital history. *History of Psychology*, **21**(4): 380–83 doi:10.1037/hop0000113.

Hélin, É. (1963). *La Démographie de Liége Aux XVIIe et XVIIIe Siécles*. (Académie Royale de Belgique. Classe Des Lettres et Des Sciences Morales et Politiques. Mémoires, Coll. in-8° 56/4). Brussels: Palais des Académies.

Klep, P. M. M. (1991). Population Estimates of Belgium, by Province (1375-1831). In Société Belge de Démographie (ed), *Historiens et Populations. Liber Amicorum Étienne Hélin*. Louvain-la-Neuve: Academia, pp. 485–507.

Paping, R. F. J. (2014). General Dutch Population development 1400-1850: cities and countryside. Alghero, Italy http://hdl.handle.net/11370/d057464a-dbb1-4d50-a217-762403c1a3e2.

Peels, R. (2019). Replicability and replication in the humanities. *Research Integrity and Peer Review*, **4**(1): 2 doi:10.1186/s41073-018-0060-4.

Peels, R. and Bouter, L. (2018a). Humanities need a replication drive too. *Nature*, **558**(7710): 372–372 doi:10.1038/d41586-018-05454-w.

Peels, R. and Bouter, L. (2018b). The possibility and desirability of replication in the humanities. *Palgrave Communications*, **4**(1): 95 doi:10.1057/s41599-018-0149-x.

Rijcke, S. de and Penders, B. (2018). Resist calls for replicability in the humanities. *Nature*, **560**(7716): 29–29 doi:10.1038/d41586-018-05845-z.

Stabel, P. (1997). *Dwarfs among Giants: The Flemish Urban Network in the Late Middle Ages*. (Studies in Urban Social, Economic and Political History of the Medieval and Modern Low Countries 8). Leuven: Garant.

Stapel, R. J. (2017). Holland rond 1500: een geografische verkenning van de *Enqueste* (1494) en *Informacie* (1514). *Holland: Historisch Tijdschrift*, **49**(4): 177–84.

Stapel, R. J. (2020). Historical Atlas of the Low Countries (1350-1800) IISH Data Collection http://hdl.handle.net/10622/PGFYTM (accessed 10 June 2020).

Tucker, A. (2017). Replication, Visualization & Tactility: Towards a Deeper Involvement of 3D Printing in Humanities Scholarship and Research. Montreal: ADHO https://dh2017.adho.org/abstracts/230/230.pdf.

UNESCO (2021). UNESCO Recommendation on Open Science *UNESCO* https://en.unesco.org/science-sustainable-future/open-science/recommendation.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, **3**(1): 160018 doi:10.1038/sdata.2016.18.

Woude, A. M. van der (1972). *Het Noorderkwartier. Een Regionaal Historisch Onderzoek in de Demografische En Economische Geschiedenis van Westelijk Nederland van de Late Middeleeuwen Tot Het Begin van de Negentiende Eeuw*. Vol. 1. 3 vols. (A.A.G. Bijdragen 16). Wageningen: Afdeling Agrarische Geschiedenis, Landbouwhogeschool.

Notes

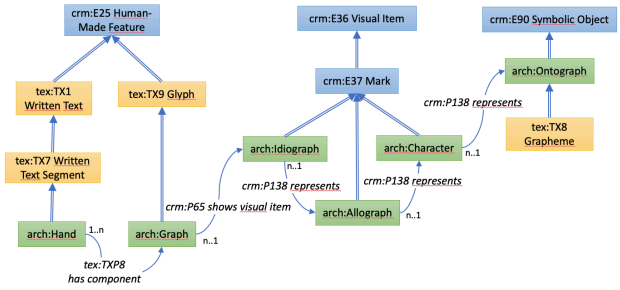
1. A second essential perquisite for replication is making it worthwhile for researchers to spend their time to replication – by opening up grant opportunities, value replication as an integral part of in scholarly careers, etc.
2. https://www.nwo.nl/en/researchprogrammes/replication-studies.

Describing Handwriting in Context

Stokes, Peter Anthony

peter.stokes@ephe.psl.eu
École Pratique des Hautes Études – Université PSL, France

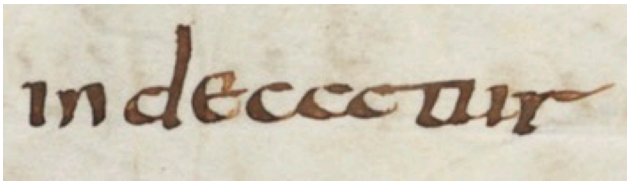
A number of papers have been given at the ADHO conferences and elsewhere regarding the modelling of writing, one of which is the Archetype model for palaeographical analysis (Diehr et al., 2018; Stokes, 2012; Stokes, 2018). As well as its implementation in the Archetype framework, this model has also been aligned to the CIDOC-CRM and CRMtex extension and then implemented in OWL (based on the Erlangen version 7.1.1), an extract of which is shown below. Although useful, this model (and indeed most others) treat the signs in written communication as distinct entities: they address the selection of a particular sign amongst a set of possible options, but they pay little or no explicit attention to the relationships between signs in a given piece of writing. In Structuralist terms, they focus primarily or exclusively on the paradigmatic aspect, but do not address the syntagmatic relations. This paper will therefore discuss some use cases and possible responses.



Extract of the Archetype model of handwriting aligned to CIDOC-CRM and CRMtex (Stokes, 2021; Stokes, forthcoming).

For Archetype, the system is designed primarily for palaeographical analysis such as the distribution of allographs by a given scribe (Brookes et al., 2015), but the choice of allographs is largely or even entirely determined by context such as the position in a word or the formality of the script (Morison, 1972), and the way in which a letter is written will often be influenced by that which immediately precedes or follows it, so an effective model should take this into account. [1]

Another important use-case is meaning and disambiguation. As Robinson has pointed out (2009: 42–43), most examples of **i** in medieval manuscripts do not comprise 'a line with a dot over it', but we recognise it anyway; however, this recognition depends in part on the linguistic context, and because we compare it with other nearby signs and adjust our expectations accordingly. This ambiguity is relatively common in practice, particularly in multigraphic contexts (for examples and discussion see e.g. Coueignoux, 1983: 61; Bugarski, 1993; Pierazzo, 2015; Stokes, 2018) and again suggests the need to model syntagmatic relationships.



Ambiguity in script: the word reads 'indecatur', but the a is indistinguishable from two cs (Pierazzo, 2015: 87). St Gallen Stiftsbibliothek MS 189, p. 76 (detail).

A third and very large case the function of signs. As Klinkenberg and Polis (2019: 68 ff.) and others have discussed, graphemic function is often relational, whether paradigmatic or syntagmatic. One example of a syntagmatic relational function is found in alphabetic systems, where letters often influence or determine the phonetic value of others. In English, for example, the pronunciation of the <a> in *hate* is partly determined by the following <e> (contrast *hat*). In some cases such as French, letters indicate the grammatical accord or function of words even when they are not pronounced (Klinkenberg and Polis, 2019: 85): contrast *il mange* and *ils mangent*. Punctuation and features such as underlining or italic also function syntagmatically, modifying the sense and/or pronunciation of that which follows and/or precedes them. Similarly, Anis (1983: 33–34) proposed graphematic supra-segmentals to describe elements in print such as italics when they are graphematically distinctive (contrast 'dating *Beowulf*' with 'dating Beowulf'). Determinants in writing-systems such as Egyptian hieroglyphics and Sumerian cuneiform also operate at this level, since the sign is usually at the start or end of a word but establishes the reading for that word, in terms of meaning and also often in sound. Finally, in

some systems such as Japanese *kanji*, the pronunciation and meaning of a sign can change dramatically depending on its context.

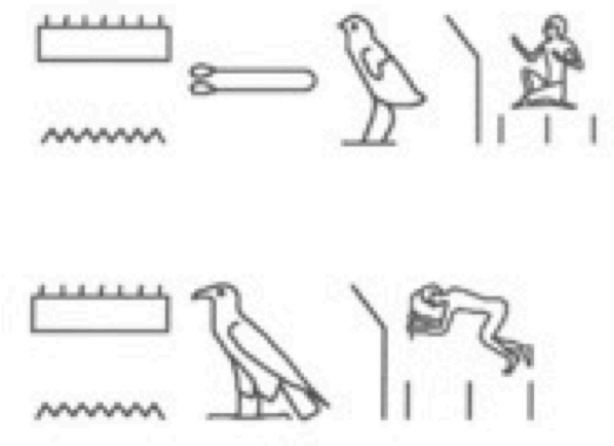


Figure 3:
Determinatives in Egyptian hieroglyphics (Klinkenberg, 2005: 178). The two words have the same pronunciation, but different determinatives ('human' and 'enemy').

Finally, the cases so far all assume linear or 'chronosyntactic' writing, but one also finds 'toposyntactic' cases where meaning is determined in part by the spatial relationship between signs (Klinkenberg, 1996; Klinkenberg, 2005: esp. 190-95). This is particularly the case in Egyptian hieroglyphs (Polis, 2018) but can be found even in alphabetic scripts, tables being an obvious example (see Pierazzo and Stokes, 2011, for some others).

The question, then, is how to model these relationships. The most obvious is through graph or network models, where relationships can be explicitly recorded. For Archetype, this would comprise different types of relationships between instances of arch:Ontograph (essentially generalised graphemes: see Stokes, 2021). For instance an arch:Ontograph *modifies phonology of* other arch:Ontographs; a determinant arch:Ontograph *modifies signification of* the other arch:Ontographs in the same word; and so on. The example of "dating *Beowulf*" is less clear, and here Anis' concept of the suprasegmental grapheme (or arch:Ontograph) seems necessary; an arch:Suprasegmental_Ontograph then groups instances of arch:Ontographs and is itself associated with a given function or other indicator. A similar approach can be used to indicate other factors, such as the language of a text or even the script (or formality of) where that is relevant. However, this raises difficulties over the distinction between the graph, the allograph, and the grapheme. CRMtex is clear that the style of script (TX10 Style) is associated with a physical instance of writing (TX7 Written Segment: Murano et al., 2020: 20–22), and this seems correct insofar as graphemes are by definition independent of details such as style. However, CRMtex does not consider allographs or idiographs, and style (and other suprasegmental markers) are at least partly also associated with these. This then suggests several types of suprasegmental, each of which specifies syntagmatic relationships between letters but at different levels of 'letter': denotative for the grapheme, connotative for the allograph and graph, topological for the graph, and so on (see also Monella, 2020a; Monella, 2020b for denotation and connontation in this context).

The conceptual model proposed here is undoubtedly complex and probably too unwieldy for most practical purposes. It is also far from complete and has its own limitations. However, it is offered as a step towards a common conceptual model for handwriting that would allow for the mapping of different projects and datasets, and, perhaps more importantly, could provide a framework for analysis and comparison of different scripts in different alphabets and writing systems. In addition, such modelling has already proven important in the refining and increasing precision in the fraught subject of palaeographical terminology and descriptions (Stokes, 2015). To paraphrase Donald Knuth (1982: 5), the process of trying to produce as complete a conceptual model as possible is surely instructive to all concerned.

Bibliography

Anis, J. (1983). Pour une graphématique autonome. *Langue française*, **59**(1): 31–44 doi:10.3406/lfr.1983.5164.

Brookes, S., Stokes, P. A., Watson, M. and Matos, D. M. D. (2015). The DigiPal Project for European Scripts and Decorations. In Conti, A., Da Rold, O. and Shaw, P. A. (eds), *Writing Europe 500–1450: Texts and Contexts*. Woodbridge: D. S. Brewer, pp. 25–58 www.jstor.org/stable/10.7722/j.ctt17mvjbg.9 (accessed 22 October 2021).

Bugarski, R. (1993). Graphic relativity and linguistic constructs. In Scholes, R. J. (ed), *Literacy and Language Analysis*. London: Routledge, pp. 5–18.

Coueignoux, P. (1983). Approche structurelle de la lettre. *Langue française*, **59**(1): 45–67 doi: 10.3406/lfr.1983.5165.

Davis, T. (2007). The Practice of Handwriting Identification. *The Library: The Transactions of the Bibliographical Society*, **8**(3): 251–76.

Diehr, F., Gronemeyer, S., Prager, C., Wagner, E., Diederichs, K., Grube, N. and Brodhun, M. (2018). Organising the Unknown: A Concept for the Sign Classification of not yet (Fully) Deciphered Writing Systems Exemplified by a Digital Sign Catalogue for Maya Hieroglyphs. *Digital Humanities 2018 Book of Abstracts*.

Mexico City: Red de Humanidades Digitales A. C. https://dh2018.adho.org/en/organising-the-unknown-a-concept-for-the-sign-classification-of-not-yet-fully-deciphered-writing-systems-exemplified-by-a-digital-sign-catalogue-for-maya-hieroglyphs/ (accessed 6 December 2021).

Klinkenberg, J.-M. (1996). *Précis de Sémiotique Générale*. De Boeck Université.

Klinkenberg, J.-M. (2005). Vers une typologie générale des fonctions de l'écriture. De la linéarité à la spatialité. *Bulletin de La Classe Des Lettres*, **1–6**: 157–96.

Klinkenberg, J.-M. and Polis, S. (2019). Les fonctions de l'écriture : un modèle général Liège https://orbi.uliege.be/handle/2268/241566 (accessed 27 November 2021).

Knuth, D. E. (1982). The Concept of a Meta-Font. *Visible Language*, **16**(1): 3–27.

Monella, P. (2020a). An Ontology for Digital Graphematics and Philology Wuppertal http://www1.unipa.it/paolo.monella/wuppertal2020/ (accessed 2 October 2021).

Monella, P. (2020b). Un'ontologia della grafematica per la filologia digitale Siena http://www1.unipa.it/paolo.monella/siena2020/ (accessed 2 October 2021).

Morison, S. (1972). *Politics and Script: Aspects of Authority and Freedom in the Development of Graeco-Latin Script from the Sixth Century B.C. to the Twentieth Century*. (Edited and Completed by N. Barker, The Lyell Lectures, 1957). Oxford: Clarendon Press.

Murano, F., Felicetti, A. and Doerr, M. (2020). *Definition of the CRMtex: An Extension of CIDOC CRM to Model Ancient Textual Entities*. Version 1.0. ICOM-CIDOC http://www.cidoc-crm.org/crmtex/sites/default/files/CRMtex_v1.0_March_2020.pdf (accessed 18 September 2021).

Pierazzo, E. (2015). *Digital Scholarly Editing: Theories, Models and Methods*. Farnham, Surrey, UK ; Burlington, VT: Ashgate.

Pierazzo, E. and Stokes, P. A. (2011). Putting the Text back into Context: A Codicological Approach to Manuscript Transcription. In Fischer, F., Fritze, C. and Vogeler, G. (eds), *Kodikologie und Paläographie im digitalen Zeitalter 2 - Codicology and Palaeography in the Digital Age 2*, vol. 3. Norderstedt: Books on Demand (BoD), pp. 397–429.

Polis, S. (2018). The Functions and Toposyntax of Ancient Egyptian Hieroglyphs: Exploring the Iconicity and Spatiality of Pictorial Graphemes. *Signata*(9): 291–363. doi: 10.4000/signata.1920.

Robinson, P. M. W. (2009). What Text Really is Not, and Why Editors have to Learn to Swim. *Literary and Linguistic Computing*, **24**(1): 41–52 doi: 10.1093/llc/fqn030.

Stokes, P. A. (2012). Modelling Medieval Handwriting: A New Approach to Digital Palaeography. In Meister, J. C., Schönert, K., Lomsché, B., Schernus, W., Schüch, L. and Stegkemper, M. (eds), *DH2012 Book of Abstracts*. Hamburg: Hamburg University Press, pp. 382–85.

Stokes, P. A. (2015). Digital approaches to paleography and book history: some challenges, present and future. *Frontiers in Digital Humanities*, **2**(5). doi: 10.3389/fdigh.2015.00005

Stokes, P. A. (2018). Modelling Multigraphism: The Digital Representation of Multiple Scripts and Alphabets. In Palau, J. G. and Russell, I. G. (eds), *Digital Humanities 2018 Book of Abstracts*. Mexico City: Red de Humanidades Digitales A. C., pp. 292–96.

Stokes, P. A. (2021). *The Archetype Ontology*. doi: 10.5281/zenodo.5771601. https://github.com/pastokes/archetype-ontology.

Stokes, P. A. (forthcoming). Aligning Archetype: Towards a Formal Model for a Transversal Palaeography. *COMSt Bulletin*.

Notes

1. For the use of 'allograph', 'idiograph' and 'graph' in this discussion, see Stokes (2012) drawing on Davis (2007).

Towards a crowdsourced linked open knowledge base of East Asian historical sources

Sturgeon, Donald
djs@dsturgeon.net
Durham University

Introduction

Substantial records of East Asian history exist in sources written in the classical Chinese language, covering important aspects of both Chinese history as well as the histories of many historical states and dynasties throughout the region, including those overlapping with the modern regions of Korea, Japan, and Vietnam. To date, substantial amounts of relevant primary source material have been digitized and transcribed, while many more materials are in the process of digitization. While digital editions are already enormously valuable to researchers, their utility can be greatly improved by semantic contextualization of aspects of their content – for example, by connecting

mentions of historical people, places, events, bureaucratic structures, time periods, and similar entities to concrete data about these entities as well as to other mentions of the same concept. This benefits human readers by enabling contextualized reading assistance and improved search functionality based on semantic data (mentions of an entity) rather than purely surface-level textual content (strings of words – in the Chinese case, strings of characters, as word boundaries are not recorded in either the premodern or modern writing system). It also facilitates quantitative analysis of important aspects of the historical record, and lays the groundwork for fully automated annotation of – and knowledge extraction from – vast corpora of premodern sources.

This paper introduces a scalable approach to the creation of a large dataset of such material, intended to provide a sustainable mechanism for annotation and knowledge base construction in a fully crowdsourced environment. While previous work (such as Simon et al 2015, and De Weerdt et al 2016) has generally used the approach of connecting mentions in a static text (which once annotated cannot easily be edited) to entities in a static knowledge base (which may occasionally be updated at intervals, but generally changes very infrequently), the work described in this paper aims to connect a dynamic text (which may require corrections at any point during or after the annotation process) to a dynamic knowledge base (which is expected to continually evolve and grow during – and as a consequence of – the process of annotation).

This work offers the following contributions: 1) the creation of a crowdsourced XML-based annotation framework building on and integrated with an existing digital library of over 5 billion characters of Chinese premodern text; [1] 2) a linked open knowledge graph of knowledge extracted from annotated texts and serialized as RDF; [2] 3) a semi-automatic approach for automatically extracting knowledge systematically from annotated texts; [3] and 4) a fully machine-readable representation of East Asian dates which avoids the previously ubiquitous Eurocentric approach of representing dates only through their conversion to Gregorian or Julian calendars – leading to direct practical benefits through use of a data model more appropriate to the task. [4]

Annotation and knowledge extraction

In this implementation, annotations are used to supplement textual content with just two pieces of information: an entity identifier in the local knowledge base, and the type of the entity referenced; any further information known about the entity is encoded separately in the knowledge base itself. [5] The only exception to this is for date references, which also encode data numerically representing literal information (often contextually provided) about the meaning of the date in its original expression – such as a year, month, and day (in the relevant historical calendar used in the text), corresponding to an offset within the specific time period represented by the linked entity, which itself refers to either a historical ruler or named era. This provides a straightforward method for encoding historical date references on their own terms, without requiring calendar conversion at the point of representation (Figure 1); instead, interpretation and calendar conversion operate as separate processes performed in real time using a comprehensive model of East Asian calendar dates, created in part using open data published in Bingenheimer et al (2016).

A purpose-designed annotation client (Figure 2) implements browser-based functionality to create and correct annotations both manually and automatically using a variety of approaches, including automatic suggestions using the current state of the knowledge base, and explicit tagging using flexible user-specified rules and lists of correspondences between patterns and entity identifiers. Annotation and knowledge graph state are updated by communication with the digital library via publicly documented APIs.

The annotation client also provides manual and semi-automatic functionality to directly augment the knowledge base itself during and after the annotation process (Figure 3). Knowledge is encoded in a structure closely following that of Wikidata, in which all information is stored as a series of *claims,* each representing a subject-verb-object sentence, each optionally associated with a series of qualifications (qualifier-object pairs) that qualify the claim. Initial work has created over 400,000 annotations of over 60,000 entities, with over 300,000 knowledge claims, covering a period of almost 3000 years of East Asian history. [6]

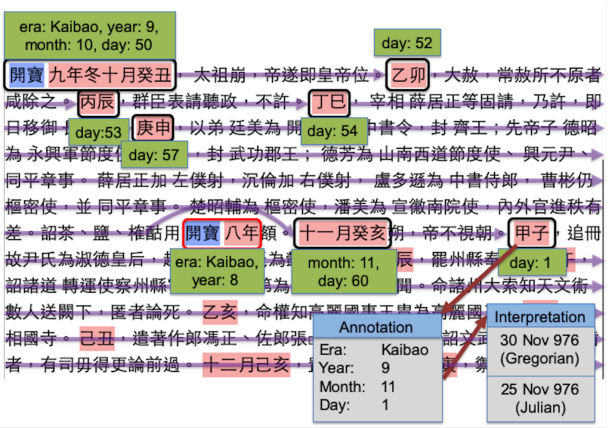


Figure 1:

Flow of contextual date information across a fragment of text. Indicated in blue boxes are explicit references to eras; pink boxes are references to dates. English text in green glosses show the literal content of the annotated text circled. The correct annotation for the last highlighted date reference (which in the text literally contains only numerical content meaning "1st in a cycle of 60") supplements this with a reference to the era entity, a specific year and month, and indicates that the numerical content refers to a day (as opposed to a year). Note that the flow of information is non-trivial, because the reference to year 8 is parenthetical and does not alter the year of the subsequent date references, which still refer to year 9.



Figure 2:
Entity annotation interface, showing linked data in the 15 th chapter of the History of the Song Dynasty (宋史).



Figure 3:
Automatic suggestions during manual knowledge claim input. An entity representing the office of 樞密使 has been suggested as the object of the verb "held-office"; a machine-readable date incorporating textual context corresponding to "明道元年十二月壬寅" (Mingdao era, year 1, month 12, day 39 [sexagenary]), and resolving to 8 January 1033 AD (Julian) has been suggested as the value for the "from-date" qualifier for this claim based on existing annotations in the text.

Bibliography

Bingenheimer, M., Hung, J.-J., Wiles, S. and Boyong, Z. (2016). Modelling East Asian Calendars in an Open Source Authority Database. *International Journal of Humanities and Arts Computing*, **10**(2). Edinburgh University Press 22 George Square, Edinburgh EH8 9LF UK: 127–44.

Simon, R., Barker, E., Isaksen, L. and Soto Cañamares, P. de (2015). Linking early geospatial documents, one place at a time: annotation of geographic documents with Recogito. *E-Perimetron*, **10**(2): 49–59.

Sturgeon, D. (2020). Digitizing Premodern Text with the Chinese Text Project. *Journal of Chinese History*, **4**(2). Cambridge University Press: 486–98 doi:10.1017/jch.2020.19.

Sturgeon, D. (2021). Chinese Text Project: A dynamic digital library of premodern Chinese. *Digital Scholarship in the Humanities*, **36**(Supplement_1): i101–12 doi:10.1093/llc/fqz046.

Weerdt, H. D., Ming-Kin, C. and Hou-Ieong, H. (2016). Chinese Empires in Comparative Perspective: A Digital Approach. *Verge: Studies in Global Asias*, **2**(2). University of Minnesota Press: 58–69 doi:10.5749/vergstudglobasia.2.2.0058.

Notes

1. https://ctext.org/instructions/annotation
2. https://ctext.org/tools/linked-open-data
3. https://ctext.org/instructions/annotation/client
4. https://ctext.org/instructions/annotation#conventions
5. Of these, the type could in principle be omitted, as it can always be inferred from the entity record. However, its inclusion directly in the annotation is practically useful for performance reasons: some common actions, such as visually indicating the presence of annotations, rely only on the type of the annotation. Additionally, unlike many attributes, in the system described, entity type always has exactly one value for a given entity.
6. https://ctext.org/tools/linked-open-data

Intertextuality in Large-Scale East Asian and Western European Corpora

Tharsen, Jeffrey

tharsen@uchicago.edu
The University of Chicago, United States of America

Gladstone, Clovis

clovisgladstone@uchicago.edu
The University of Chicago, United States of America

Intertextuality has been a significant concern of scholarly communities around the world for centuries; fields like *Redaktionsgeschichte* in Germany and *jiaokanxue* 校勘學 in China have long provided evidence-based foundations for debates on the relationships between works, editions and authors. With the advent of digital texts and computational tools, new avenues for research into intertextuality have recently emerged. In this presentation, we will discuss our initial results from the TextPAIR framework, a language-agnostic open-source unsupervised approach to detecting "text reuse" in any language or script, and discuss future avenues for algorithmically-based research into relationships between textual communities, traditions and sources.

Like the vast majority of data-mining or machine-learning tools, the results provided by TextPAIR are heavily impacted by the input data. We show that TextPAIR's multilingual capabilities are made possible first by the flexible approach to text representation, that is how texts are preprocessed and transformed prior to being handed over to the matching algorithm. Because each language is different, be it structurally or culturally, different aspects of language may be retained from one language to the other. In other words, the goal of the preprocessing stage is to bring forward the ideal features needed for maximizing the quality and yield of the text-reuse detection algorithm. We will thus describe this process of parameterizing the preprocessing steps in order to obtain a text representation that is appropriate for the language of text collections(s) being analyzed.

Through the analysis of text reuses found within large-scale corpora in different languages, with a focus on Chinese, Japanese, French and English, we highlight how the nature of the reuses reflect particular linguistic and cultural traditions, especially in the vast number of cases where the reuse is not identical to the source. For example, the preliminary results of our research indicates that reuses detected in European languages tend to highlight orthographic variations, while reuses in Chinese (as the language employs a logographic script) evidence replacement and editorial decision-making at the level of the individual character, and reuses in our Japanese *Aozora Bunko* 青空文庫 corpus demonstrate how authors (and/or editors) chose to employ hiragana versus kanji when citing the "same" source. As we continue developing the toolkit we are looking to incorporate new large-scale corpora in Russian, German, Spanish, Arabic, Hindi and Urdu, with the goal of determining if the reuses in those datasets hew more towards what we have found in European contexts or are more similar to the types of reuses in our East Asian corpora.

We then move on to describing how our approach to intertextuality builds upon previous efforts such as networks of identifiable citations (Long and So, 2013) or general correspondences between intellectual traditions (Kristieva, 1980) by providing an end-to-end system for the unsupervised detection and elucidation of text reuse in all its forms, from "identical" (and lightly modified) passages to various types of allusions to conceptual parallels between and among works. Thanks to the scalability of TextPAIR, we are able to combine the raw output of many thousands of text reuses with network-based visualization techniques in order to discern previously unseen patterns within particular intertextual traditions, such as communities of works and authors that tend to rely on similar sources even though they may not borrow directly from one another. This can take many forms; for example: a group of authors who reuse the same texts to strengthen their arguments, as in a study we conducted on the uses of Lucretius in 18 th century England (Cooney and Gladstone, 2020), or which original sources are unique to versus shared between the twenty-four Chinese official histories (Tharsen and Gladstone, 2020).

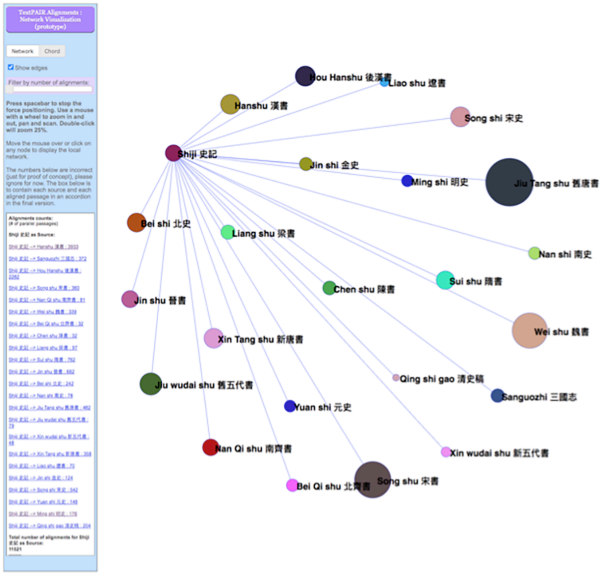
Finally, we explore the wealth of new approaches these methods make possible: detection of correspondences across multiple languages and through various intellectual traditions by leveraging advances in automatic translations made possible by deep learning methods, new ways to map the development of ideas and concepts over the *longue durée* provided by new matching methods within the TextPAIR framework, and insights into the sources of many of our most classic works, long obscured by time, space and/or lack of prestige.

Screenshots and Links:

The 二十四史 *"Twenty-four Chinese Histories" (322,000 text reuses; 90 BCE to 1927)*

TextPair UI: https://anomander.uchicago.edu/text-pair/histories/

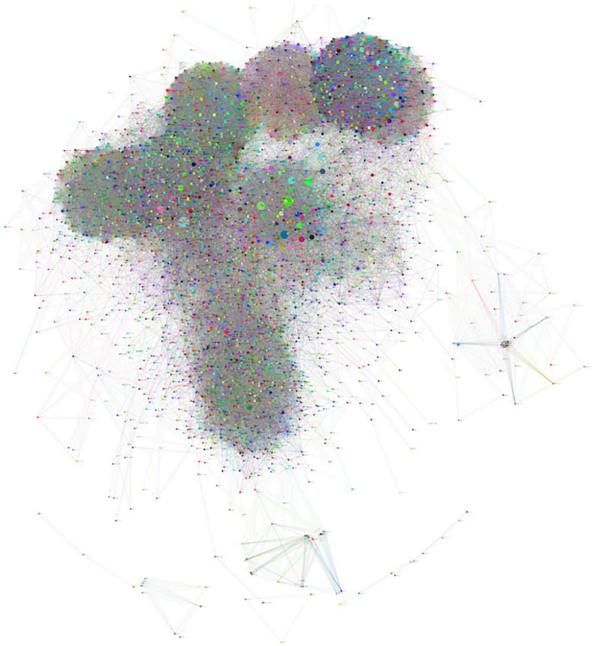
TextPair Viewer: https://users.rcc.uchicago.edu/~jcarlsen/TPV/TPV_histories/



The 二十四史 *"Twenty-four Chinese Histories" divided by chapter/* juan 卷

TextPair UI: https://anomander.uchicago.edu/text-pair/histories_juan_flex/

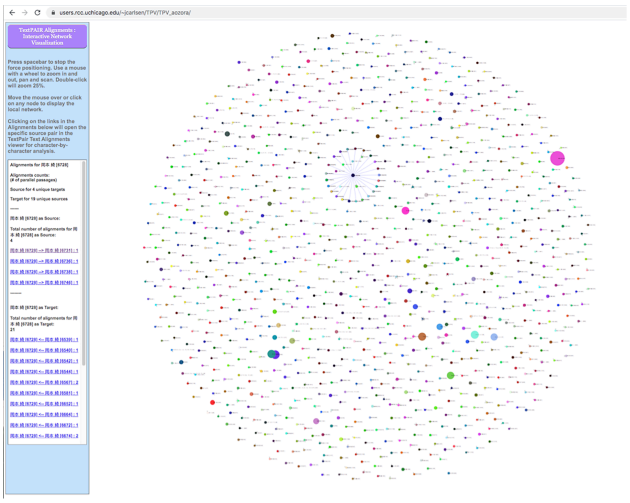
TextPair Viewer: https://users.rcc.uchicago.edu/~jcarlsen/TPV/TPV_histories_juan_flex/



Text reuses in the Aozora Bunko **青空文庫** *(over 15,000 works; 1,219 text reuses)*

TextPair UI: https://anomander.uchicago.edu/text-pair/aozora/

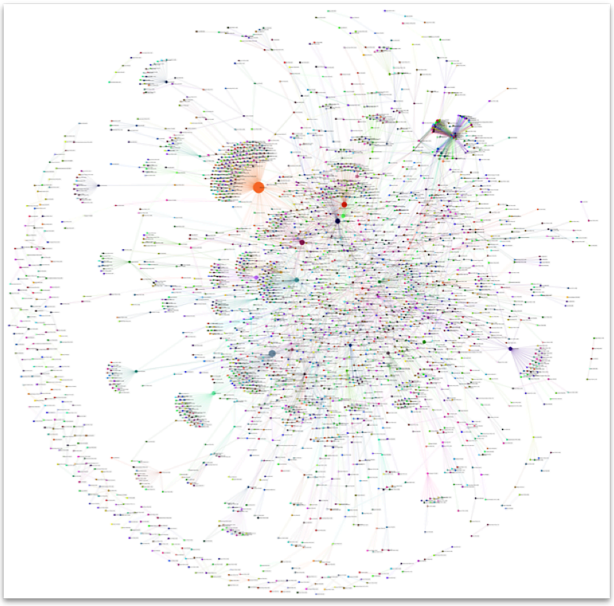
TextPair Viewer: https://users.rcc.uchicago.edu/~jcarlsen/TPV/TPV_aozora/

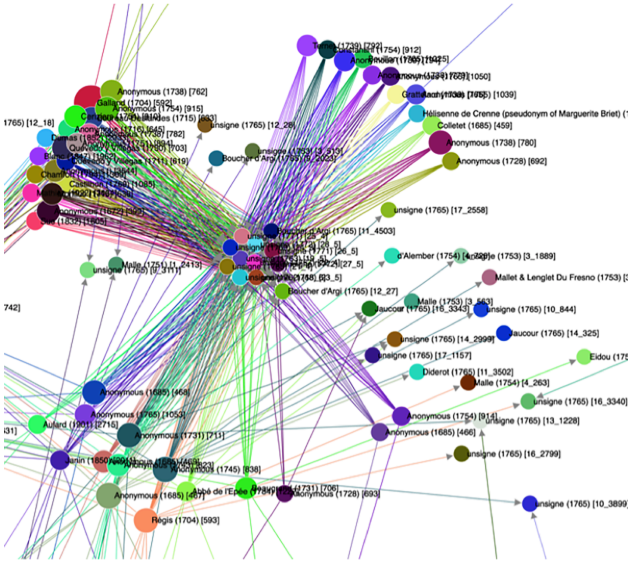


Diderot's Encyclopédie *and 18 th -century French literature (over 7,000 text reuses)*

TextPair UI: https://anomander.uchicago.edu/text-pair/encyc/

TextPair Viewer: https://users.rcc.uchicago.edu/~jcarlsen/TPV/TPV_frantext_encyc/





French Literature & Diderot's Encyclopédie *: Closeup of one cluster*

Bibliography

Cooney, Charles and Gladstone, Clovis. (2020) "Opening New Paths for Scholarship: Algorithms to Track Text Reuse in ECCO", in *Digitizing Enlightenment: Digital Humanities and the Transformation of Eighteenth-Century Studies*, Simon Burrows & Glenn Roe ed., Oxford University Studies in the Enlightenment, Voltaire Foundation in association with Liverpool University Press.

Kristeva, Julia. (1980) *Desire in Language: A Semiotic Approach to Literature and Art*. New York: Columbia University Press.

Long and So (2013) "Network Analysis and the Sociology of Modernism." *Boundary* 40, no 2.

Olsen, M., Horton, R., & Roe, G. (2011). "Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections." *Digital Studies/le Champ Numérique*, 2(1).

Smith, David A.; Cordell, Ryan; Dillon, Elizabeth Maddock (2013). "Infectious Texts: Modeling Text Reuse in Nineteenth-Century Newspapers," *Proceedings of the Workshop on Big Humanities*, IEEE Computer Society Press.

Sturgeon, Donald (2018). "Unsupervised Identification of Text Reuse in Early Chinese Literature," *Digital Scholarship in the Humanities* 33, no. 3.

Tharsen, Jeffrey, and Gladstone, Clovis (2020). "Using Philologic For Digital Textual and Intertextual Analyses of the Twenty-Four Chinese Histories 二十四史." *Journal of Chinese History* 4, no. 2.

Vierthaler, Paul and Gelin, Mees (2019). "A BLAST-based, Language-agnostic TextReuse Algorithm with a MARKUS Implementation and Sequence Alignment Optimized for Large Chinese Corpora," *Journal of Cultural Analytics*.

Connecting Digital Systems for whom and by whom? Taking Stock of the Digital Research Infrastructures in China

Tsui, Lik Hang
lhtsui@cityu.edu.hk
City University of Hong Kong

Chen, Jing
cjchen@nju.edu.cn
Nanjing University

The term "cyberinfrastructure" is derived from research in the natural sciences. In a 2003 NSF report, cyberinfrastructure was described as "an idea that has emerged… from some basic technological realities," which "have led researchers to envision a tightly integrated, planet-wide grid of computing, information, networking and sensor resources" (Hart, 2003). As such, a cyberinfrastructure is akin to a new knowledge environment and an ecology of virtually connected organizational resources, rather than research materials only. Later, the ACLS specified the role of cyberinfrastructure in humanities research (Courant et al., 2006). More than 15 years has passed now; it seems that the conception of a cyberinfrastructure in digital humanities still fits such criteria in the American scholarly community. Other strategies have also been devised in Europe, UK, and Australia etc. However, is the "cyberinfrastructure" idea readily applicable to research communities in non-Western contexts? How should this issue be evaluated and who should serve as the target users and participants of this infrastructure? Based on the DH2022 theme, this study will zoom in on an Asian context, especially the Chinese-speaking world.

Since the end of the 2000s decade, there have been increasing discussions advocating or taking part in the building of cyberinfrastructure in China, especially to link various resources and systems for digital research (刘炜, et al., 2016; 王宏甦、徐力恒、包弼德, 2020; 刘炜, 2020). Chinese researchers have been developing digital projects for decades, at least from the 1980s (徐力恒, 2020; Chen & Tsui, 2022). Substantial preparation is already in place for an infrastructure. Our study critically analyzes such efforts

in mainland China and also serves to globalize and diversify the discussions on digital research infrastructures. This takes the form of an evaluation of the current state of the field and an outline of some theoretical provocations in order to achieve an understanding of the epistemological and socio-technical differences that arise when building research infrastructures in diverse socio-cultural environments.

Digital projects are developed in the region so that users dealing with data in Chinese could upload them to online platforms, as well as harness tools to analyze and visualize them in an online, mostly open and sharable environment. Examples of platforms that have a presence among users in mainland China include DocuSky, Jihe Net 籍合网, MARKUS, Shanghai Library's Open Data Platform, Zhejiang University's Academic Map Publishing Platform, and several others (Tsui, 2020). By the 2020s, there is already substantial preparation for a digital infrastructure for humanities research in China.

Our intervention poses two questions about the digital infrastructures in China that are taking shape. First of all, for whom is the infrastructure in China built? Since the drive that we see in this is mostly academic, the incentives are primarily about producing academic outputs, such as scholarly articles. The role of community projects in the digital research infrastructure is still largely absent. Community involvement is also seriously lacking in academic projects, even when the aforementioned platforms are already in place and operative (陈静, 2018). The infrastructure, even if it is for supporting academic research, should not be confined to the academe. Since it should inherently be an "infrastructural" and basic system, it should be open and should not incur too much costs to access for the participants and stakeholders.

A second question is by whom this cyberinfrastructure should be built. Even though some standard procedures and workflows for such an infrastructure are already under discussion, there are still challenges in the development of digital humanities in China (Zhu & Zhang, 2020). For instance, there are barriers for the streamlined text mining of Chinese data, including in word segmentation. Humanists also face various problems when digitizing their materials. These are challenges that a cyberinfrastructure could help tackle. Who then should be building a broad-purposed cyberinfrastructure that serves these general but foundational and much-needed functions for Chinese(-originating and Chinese-language) data? Ideally, the infrastructure should also include utilities, software, and algorithms that aid such digital humanities endeavors in the Chinese world. Most of the actors engaging in the building of an infrastructure for digital humanities in China are from the fields of engineering, computer science, and information technology, coming from libraries, university digital humanities centers, and database companies. The participation of humanists, artists, product designers, user experience designers, and community members are still relatively lacking. A collaborative model in which these actors come together to work on this is much needed.

Bibliography

Chen, J. and Tsui, L. H. (2022). Debating and Developing Digital Humanities in China: New or Old? In Fiormonte, D., Chaudhuri, S. and Ricaurte, P. (eds), Global Debates in the Digital Humanities. Minneapolis: University of Minnesota Press, 2022, pp. 71-86.

Courant, P. N., et al. (2006). Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences. http://acls.org/uploadedFiles/Publications/Programs/Our_Cultural_Commonwealth.pdf (accessed 21 April 2021).

Hart, D. (2003). Cyberinfrastructure: A Special Report. https://www.nsf.gov/news/special_reports/cyber/Cyberinfrastructure%20_NSF.pdf (accessed 21 April 2021).

Tsui, L. H. (2020). Charting the Emergence of the Digital Humanities in China. In Chan, K. Y. and Lau G. C. S. (eds.), Chinese Culture in the 21st Century and its Global Dimensions: Comparative and Interdisciplinary Perspectives. Singapore: Springer, 2020, pp. 203-216.

Zhu, B. and Zhang, J. (2020). Digital Humanities Cyberinfrastructure for Ancient China Studies: Past, Present, and Future, Library Trends, 69 (1): 319-333.

包弼德、夏翠娟、王宏甦 (2018). 数字人文与中国研究的网路基础设施建设, 图书馆杂志, 37 (11): 18-25.

陈静 (2018). 当下中国 "数字人文" 研究状况及意义, 山东社会科学, 7: 61-65.

刘炜, et al. (2016). 面向人文研究的国家数据基础设施建设, 中国图书馆学报, 42 (5): 29-39.

刘炜 (2020). 作为数字人文基础设施的图书馆: 从不可或缺到无可替代, 图书馆论坛, 40 (5): 1-2.

王宏甦、徐力恒、包弼德 (2020). 用于中国历史研究的网路基础设施：对相关探索的建议和展望, 数位典藏与数位人文, 6: 1-35.

徐力恒 (2020). 华文学界的数位人文探索：一种 "史前史" 的观察角度, 中国文哲研究通讯, 30 (2): 107-127.

Object constitution in digital collections: An Ethnomethodological View

Türkoglu, Enes

enes.tuerkoglu@uni-koeln.de
University of Wuppertal, Germany

Mertgens, Andreas

mertgens@uni-wuppertal.de
University of Wuppertal, Germany

Libraries, archives, museums, collections or other similar memory and research institutions collect, preserve, organize, exhibit and grant access to culturally, informatively and historically significant objects from heterogeneous domains, because these objects act as proxies for "epistemic things" of real word phenomenons. As a result of these processes, objects become things of knowledge. The systematic "Wirklichkeitserschließung" (roughly: indexing of reality) through the organization of the objects can be seen as a process of intellectual measurement, which in turn is a product of specific context-dependent cultural grammar (Marx 2020). Therefore, this form of "measurement" can not claim a status of objectivity, and it can not fully exhaust the significance of the objects. Especially, the loss of context by separation from the initial cultural practices and modes of circulation cannot be counteracted solely through terminological or functional categorization. This phenomenon is prominently discussed in the museum context on the axis of de- and recontextualization (Jones 1992)(Malraux 1960). Similar concerns are raised in archival discourse, in which document collections are disassembled, textual and visual material are separated, disregarding their respective context. (Darms 2009)

Through digitization efforts of the collected objects, these troubles are deepened. With digitization, it can be observed that the objects go through another layer of decontextualization and intellectual appropriation. On the one hand the objects get separated from their initial materiality as they get digitized, on the other hand decisions are being made that impact what and how contextual information is recorded and represented (Beaudoin 2012). Moreover, this layer constitutes the objects themselves akin to the previous constitution by the activities of memory institutions. The objects are decontextualized from their social circulation and cultural practices as they are collected and preserved, then they are made digitally operational and with that they get decontextualized and reconstituted again. The activity of making objects digitally operational and representable is discussed intellectually in DH with the terms models and modelling (Ciula et al. 2018), and their complex relationship with the target object explained through the notion of "creative process" to underline the productive nature of the activity rather than reproductive.

As argued, initial institutional activities convey "Wirklichkeitserschließung"; comparably, digitization conveys indexing and exploration of objects as epistemic things derived from real world phenomenons. In his work "how reproductive is reproduction?" Björk investigates how the informative configuration of the digital archival documents is situationally conditioned. Björk argues that the institutional ideals can intervene in the relationship of objects and their potential informative capacity as they get digitized and made accessible. Similar notions with slightly different perspectives can be observed.

Under these conditions there is a need for discourse about the practices of digitization, collection and organization that constitute objects of interest and with that our understanding of reality. Within the Collaborative Research Center "Media of Cooperation" the opportunity to carefully scrutinize these practices arose when confronted with the task of digitizing papers from the estate of Harold Garfinkel, regarded as the pioneer of ethnomethodology and was known for his non-conventional methodologies. The digitization of his papers allows for an interesting "hybrid study" which ethnomethodology utilizes "to merge ethnomethodological studies with investigative topics treated within the settings being studied" (Ikeya 2020: 23). With this approach, a practice can be utilized, while the practices themselves get studied. To great advantage for the project, Garfinkel already worked on a theory of information and specifically on a mechanical information retrieval system *Zatocoding* as hybrid study. His insights are helpful to understand formal information practices and give clues for a conception which can take these practices meaningfully further. At its core, his theory of information describes that "'information is constituted' — not just interpreted 'or symbolically represented and exchanged but actually constituted as information by the social (cooperatively ordered) aspects of the situated social orders in which it occurs.'" (qtd. in Ekbia 2009).

In the practices of DH projects, this theory is reflected in constant self-scrutiny in regard to object-constitution. At each stage of the project - from digitization, metadata collection, indexing, information-modeling and presentation - an attempt is made to question conventional or best-practice approaches towards object constitution and to allow for differing dimensions of representation to co-exist. This includes a digitization effort focussed strictly on the material findings and on recording the structure and material context of the boxes, folders and documents within the collection. Subsequently, a more conventional finding-aid was created, which necessarily contextualized and restructured the documents into units from an archival point of view. The collection includes material that could be read as the genesis of an unpublished academic work. With this perspective, the documents would be contextualized as part of a genetic edition not based on their material findings but as academic artefacts, witnesses of research practices and

the evolution of ideas that eventually led to drafts of a work. Another approach is to structure the collection according to Garfinkel's use of language, i.e. identifying terms and phrases within and across documents.

Due to the experimental nature of the project, these varying approaches are followed concurrently with constant self-observation and documentation of the limitations and affordances of each practice. One aspect that emerged is that despite the prevalence of models and best-practices to guide such a digitization project, many decisions relied heavily on ad-hoc decisions made by individuals. This echoes Garfinkel's understanding in observing that ad-hoc decisions are an invariable and essential part of any practice and who claimed that "treating ad hoc features as a nuisance to complaining that if the walls of a building were only gotten out of the way that one could better see what was keeping the roof up" (qtd. in Caron 2013). This multivariate, self-observative, bias-conscious project configuration transposes concepts of ethnomethodology into the field and can open up discussions within the DH community. Similar to the question of how information is not just interpreted but constituted, so are the objects in Digital Humanities projects and with every "best"-practice and every standard we employ, we change the objects we constitute, preserve and present to the public.

Bibliography

Beaudoin, J. E. (2012). Context and Its Role in the Digital Preservation of Cultural Objects. *D-Lib Magazine* 18 (11/12). doi.org/10.1045/november2012-beaudoin1.

Björk, L. (2015). How Reproductive Is a Reproduction?: Digital Transmission of Textbased Documents. Borås: University of Borås.

Caron, C. O. (2013). *Reflexivity at Work: Making Sense of Mannheim's, Garfinkel's, Gouldner's, and Bourdieu's Sociology*. Canadian Theses. Library and Archives Canada.

Ciula, A., Eide, Ø., Marras, C., and Sahle, P. (2018). Models and Modelling between Digital and Humanities. Remarks from a Multidisciplinary Perspective. *Historical Social Research / Historische Sozialforschung Vol. 43* No. 4:. doi.org/10.12759/HSR.43.2018.4.343-361.

Darms, L. (2009). The Archival Object: A Memoir of Disintegration. *Archivaria* 67 (July), pp. 143-55. archivaria.ca/index.php/archivaria/article/view/13212.

Ekbia, H. (2009). Information in Action: A Situated View. *Proceedings of the American Society for Information Science and Technology* 46 (1): 1–11. doi.org/10.1002/meet.2009.1450460233.

Ikeya, N. (2020). Hybridity of hybrid studies of work: Examination of informing practitioners in practice. *Ethnographic Studies,* 17: 22-40. https://doi.org/10.5281/ZENODO.4050533.

Jones, P. (1992). Museums and the Meanings of Their Contents. *New Literary History* 23 (4): 911. doi.org/10.2307/469177.

Malraux, A. (1960). *Stimmen der Stille*. Deutsche Buch-Gemeinschaft. pp. 10-12

Marx, P. W. (2020). Maßgabe der Dinge. In *Dokumente, Pläne, Traumreste: 100 Jahre Theaterwissenschaftliche Sammlung Köln*, pp. 16-13. Berlin: Alexander Verlag.

Schweibenz, W. (2020). "Wenn das Ding digital ist …" in *Objekte im Netz.* eds. *Udo Andraschke und Sarah Wagner.* transcript. pp. 15 - 28. https://doi.org/10.1515/9783839455715-002

Using Temporal Information on Topic Mining

Uno, Takeaki
uno@nii.jp
National Institute of Informatics, Japan

Kobayashi, Ryota
r-koba@edu.k.u-tokyo.ac.jp
The University of Tokyo, Japan

Takedomi, Yuka
yuka_takedomi@nii.ac.jp
National Institute of Informatics, Japan

Hashimoto, Takako
takako@cuc.ac.jp
Chiba University of Commerce

Introduction

Analyzing social media data (e.g., Twitter and Reddit) is essential to characterize social and political problems and understand them. Practically it is impossible to read all the social media posts and digest them manually. Thus, we have to extract underlying topics from the text automatically.

Clustering is a suitable tool for extracting topics from social media data. However, it is still challenging to discover interpretable topics from a corpus of noisy and short texts, including social media posts. A limitation of the existing clustering methods, such as topic models (e.g., LDA [1]), is that they extract topics based on only

word information (e.g., word frequency in the post). It is a notable feature that the social media data is the streaming data, i.e., the data consists of the text and the timestamp (the posted times). In contrast, it is not apparent how to utilize temporal information to find interpretable topics. For instance, we may be able to classify the tweets based on only the timestamps. However, this method would not find interpretable topics because many users post various topics of tweets simultaneously. Here we propose a clustering algorithm that utilizes the word and temporal information of the posts.

Proposed Clustering Algorithm

We propose a two-stage clustering algorithm that discovers course-grained topics by leveraging textual and temporal information [2]. Suppose that a huge volume of Twitter data, with text bodies (tweets) and timestamps (posted times), is available for us.

In the first stage, we extract fine-grained topics (micro-clusters) by clustering the word co-occurrence graph of the tweets. We used the Data Polishing algorithm [3,4] to obtain the micro-clusters, that is, the tweets sharing a fine-grained topic. Data polishing algorithm was modified to incorporate the bias of the tweet data due to many retweets. Data polishing algorithm was modified to incorporate the bias of the tweet data due to many retweets. This modification improved the computational efficiency of the algorithm significantly. The temporal pattern of a micro-cluster is obtained by calculating the posting frequency of all the tweets in a micro-cluster. In the second stage, we discover the coarse-grained topics (e.g., a reaction to breaking news) by clustering the time series of the micro-clusters. We used K-Spectral Centroid (K-SC) clustering [5] to obtain the cluster of micro-clusters sharing a similar temporal pattern. K-SC algorithm is an extension of K-means, which improves the robustness to scaling and shifting. If the fine-grained topics originate from an item of news, we can expect their temporal pattern should be similar. This is the motivation for focusing on the temporal pattern to find course-grained topics.

First, the proposed algorithm was applied to large-scale Twitter data related to the "COVID-19 vaccine" [1]. We discover three types of coarse-grained topics from the Twitter data (26 million tweets): A. Reaction to the news (8 topics), B. Reaction to the tweets (2 topics), and C. Others (2 topics). While topics A and B exhibit a clear peak in their temporal pattern, topic C does not exhibit peaks: it shows a persistent activity. This topic comprises rumors, fake news, alerts for fake news, jokes, tips, etc., which are continuously posted on Twitter. Notably, the data polishing method can extract even such a less popular topic.

Next, we evaluate the computational efficacy of the proposed algorithm; we ask how fast the algorithm process the data? We compare the proposed algorithm with five existing algorithms for finding topics from tweets: LDA [1], K-means, MeanShift, Agglomerative clustering, and Data Polishing [3, 4]. We found that the proposed algorithm is much faster than the existing methods. For instance, the proposed algorithm is more than 300 times faster than LDA with a data set of 100,000 tweets.

Bibliography

Blei, D. M. (2012), Probabilistic topic models. Communications of the ACM, 55:77–84

Hashimoto, T., et al. (2021), Two-stage Clustering Method for Discovering People's Perceptions: A Case Study of the COVID-19 Vaccine from Twitter, IEEE BigData 2021.

Hashimoto, T., et al. (2021), Analyzing Temporal Patterns of Topic Diversity using Graph Clustering, The Journal of Supercomputing, 77:4375-4388

Uno, T., et al. (2017), Micro-clustering by Data Polishing. IEEE BigData 2017, 1012-1018

Yang, J. and Leskovec, J. (2011), Patterns of temporal variation in online media. In Proceedings of the fourth ACM international conference on Web Search and Data Mining, pages 177–186

Reducing Redundancy Bias in Digital Library Collections

VandenBosch, Adrienne
adrienne.vandenbosch@du.edu
University of Denver, United States of America

Organisciak, Peter
peter.organisciak@du.edu
University of Denver, United States of America

Matusiak, Krystyna K
krystyna.matusiak@du.edu
University of Denver, United States of America

The continued growth of digital libraries as primary sources for text analysis makes the issues created by duplicating and repeated texts in these collections even more apparent. The Similarities and Duplication in Digital Libraries project (SaDDL) identifies complex relationships

of digitized works in large-scale digital libraries using content-based machine learning approaches. It is unique in its approach to relationship classification: where such work is traditionally limited to metadata-based inferences, the availability of book content, collected through the HathiTrust Research Center's Extracted Features dataset (Jett et al., 2020; Organisciak et al., 2017), enabled a content-based approach for comparing book by the language used within them. It was evaluated on 8.7 million English-language works in the HathiTrust (HT) collection. Tagging these relationships helps identify duplicate and partially overlapping works in digital libraries more accurately, allowing for scholars studying historic collections using computational text mining to work with less-biased collections. This paper introduces the challenges of working with collections of such scale, the machine learning methods employed to address these challenges, and the overall implications of the project. SaDDL's outcomes include datasets that digital humanists can use to better identify works of interest and their variants in the HathiTrust and overlapping collections, as well as recommendations for most representative copies of duplicate works.

The pursuit of content-based corpus analysis is longstanding in the study of culture at scale (Manovich, 2018), analysis of texts and literature at scale (Moretti, 2013), and the study of potential patterns from corpus-wide analysis (Michel et al., 2011). The contents of large-scale digital libraries provide incredible opportunities for study in the computational examination of historical texts to identify patterns within society and culture. Duplicate and repeating text within collections can significantly impact the effectiveness of semantic models and introduce unintentional bias in analysis (Schofield et al., 2017). With the ability to accurately identify and account for duplications within corpora, these biases can be addressed in analysis and machine learning.

Outputs

Duplication in bibliographic materials is not a trivial issue, as both content and format often change, evolve, or are reconfigured. Items identified as identical works, different versions or editions, and those with whole-part relationships (i.e., anthologies or multi-part works) are all distinguished in the SaDDL dataset. In addition to tags identifying whole or partial relationships between published books, SaDDL also suggests the best representation of a work from multiple versions. The best representation of a work was determined based on the real work that most resembles an averaged view of all the copies. Recommendations of different-but-similar books are also included, trained up from an alignment of HathiTrust works with Goodreads data (Wan & McAuley, 2018; Wan et al., 2019).

The web application of the dataset is available at https://saddl.du.edu and provides an accessible interface for browsing on the volume, work, and manifestation levels. As shown in Figure 1, the web interface displays the item metadata, HathiTrust ID number, and metadata for items identified as the same work. Following the initial item information are the additional types of identified relationships grouped by type, like items that are identified as the same book but different printings or versions, as seen in Figure 1 under *Other Books With this Work*.



Figure 1:
The History of Sir Charles Grandison *in the SaDDL web interface*

Recommendations of related works are also included on the website. The website gives context to the data, helping to ensure informed use in future applications, and offers downloads of JSON files for individual books, as shown in Figure 2, and links to the full dataset. Project code and instructions for accessing the dataset are made available at the University of Denver's Massive Text Lab Github organization. Future work will deduplicate the corpus and further scholarly work over large-scale digital libraries, with implications for use across other collections.

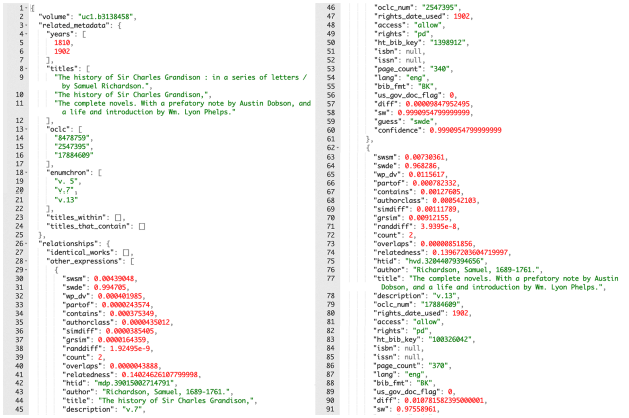


Figure 2:
A portion of the JSON file for The History of Sir Charles Grandison *from the SaDDL web interface*

Approach

To fully realize the completed SaDDL dataset, books in the HathiTrust EF Dataset were modeled using GloVe word embedding models (Pennington et al., 2014). These were chosen over newer contextual language models, such as BERT (Devlin et al., 2018) for performance and access reasons. First, their language space is linear, allowing for embeddings to be compared directly and quickly with geometric distance metrics. This choice was further motivated by the scale considerations – a large corpus comprised of texts that are much longer than typically handled in text mining applications – and the fact that the book representations that the EF Dataset provides publicly are in bag-of-word format. For document representation, book in the EF Dataset were split into similar-length chunks and projected to vectors, resulting in approximately 108 million book chunks.

To account for scale, a two-pass approach was employed for relationship evaluation: first an approximate nearest neighbor process to narrow the comparison space to candidate relationships, then a deep neural network classifier to perform relationship tagging. The SaDDL relationship classifier for full book-to-book comparison works on two inputs: a chunk-to-chunk similarity matrix, which is parsed using a convolutional neural network, and a difference vector derived from the overall book-level embeddings. Using a one-hot encoded classifier, class inferences were trained on relationships known from cataloguing metadata as well as artificially generated training examples.

Conclusion

The SaDDL dataset offers scholars a way to pursue computational text analysis over massive digital library collections while avoiding the challenges of repeating or closely redundant text. Further, it offers recommendations for choosing the most appropriate copy of a work, as well as topically-related book recommendations which may be used for developing domain-specific subcorpora. To aid in scholarly use of this data, the SaDDL dataset is accessible both in a raw format and accessible, easy-to-use website.

The present work demonstrates these products and their value for future scholars. Further, the methods for creating this work are described, which demonstrate a use of deep neural networks that is more effective than traditional feature extraction and classification workflows. The developed methods allow relationships between works that are more complicated than exact duplicates to be identified from content and can be applied to other large-scale collections. De-duplication that accounts for the complexities of publishing history and practice can be applied to all digital libraries to reduce bias and improve the results of computational analysis of historical texts.

Acknowledgements

This work was funded by IMLS #LG-86-18-0061-18.

Teaching Digital Scholarly Editing North and South Through Minimal Computing

Viglianti, Raffaele
rviglian@umd.edu
University of Maryland, United States of America

Del Rio Riande, Gimena
gdelrio@conicet.gov.ar
Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina

Abstract

In this paper, we present our reflections on whether minimal computing as a practice can extend beyond “computing done under some technological constraints” (Minimal Computing Working Group, 2021)

to serving as a common ground between different digital humanities research and pedagogical dynamics between the so called Global North and South. We explore this question by commenting on our experience in developing and teaching a bilingual undergraduate course to students enrolled from both the University of Maryland, College Park in the United States and Universidad del Salvador in Buenos Aires, Argentina. The class has so far undergone two iterations (2020 - 2021) and has introduced students to global digital humanities, digital publishing and textual scholarship of bilingual Spanish and English texts, and presented minimal computing applied to Digital Scholarly Editions (DSEs) as a shared set of values including: use of open technologies, ownership of data and code, and reduction in computing infrastructure.

Scholarly Editing Through an Open and Global Lens

The debate over a "global digital humanities" resulted in a considerable shift in 2013 when the Global Outlook Digital Humanities (GO::DH) Special Interest Group of the Alliance for the Digital Humanities Organizations (ADHO) was founded. Nonetheless, despite the benefits that we could expect from a global digital humanities, it is crucial to remember that the concept of the "global" is complex and even contradictory, especially when related to technology in all its facets, including software architecture, infrastructure as hardware, and infrastructure as long-term preservation of software.

While not all textual scholars might rely on the same definition for DSEs, they recognize their features and uses (Sahle, 2016). Free, open standards such as the ones developed by the Text Encoding Initiative (TEI), along with eXtensible Markup Language (XML) technologies, such as eXtensible Stylesheet Language Transformations (XSLT) and XQuery, and dedicated software have characterized the digital editing field. The scholarly editions themselves, however, haven't always been successful as open products of research. While it is common practice to make TEI data publicly available, the debate over how DSEs need to be structured to be truly "open" is still ongoing and best practices have yet to be established. DSEs require substantial infrastructure and advanced technical skills, while diverse needs, capacities, priorities, languages, and academic traditions may require different features at a global scale. With that in mind, how can DSEs become global? How can DSEs be minimal?

We propose minimal computing as a shared set of values such as the use of open technologies, ownership of data and code, and reduction in computing infrastructure. A minimal computing approach can moreover contribute to strengthening the bonds between the digital humanities community to open source software and hardware (OSSH) or other grassroots innovation movements that in many different regions, like Latin America or Africa, have been leading or co-creating open science projects inside academia (Arancio, 2021). In developing our syllabus, we focused on the following question: Could minimal computing provide a set of shared principles and technologies to empower students and scholars to work autonomously on and have more control over the future of their own projects? We further considered: What if minimal computing extended beyond "computing done under some technological constraints" by standing at the core of a global Digital Humanities commons, overcoming notions such as center and periphery, North and South? and more broadly: Could minimal computing serve as a common ground for Northern and Southern digital humanists?

Teaching Digital Publishing with Minimal Computing

In December 2019, we proposed a course titled Digital Publishing with Minimal Computing: Humanities at a Global Scale to the Global Classroom Initiative (GCI) program at the University of Maryland. This program offers support for the development of courses to be taught in collaboration with a higher education institution outside of the United States, with the goal of establishing courses that expose students to work that is cross-cultural, project-based, and virtual; the GCI argues that these courses mirror the work students will encounter throughout their lives. While this outcome is somewhat dependent on the students' career choices and opportunities, it is evident that "globalization shrinks the world, bringing a wider range of cultures into closer contact than ever before" (UNESCO, 2013). Thus, preparing students to participate in a globalized world is a worthwhile goal, particularly if this can be done in a way that fosters intercultural competences.

The course has received funding for at least three iterations between 2020 and 2022, with a blend of online and in-person learning. It is centered around a group project in which students collaborate virtually to create a bilingual (Spanish and English) digital edition of a multilingual colonial era text, a travelogue written by a Basque trader called Acarette Du Biscay, *An Account of a Voyage up the River de la Plata*, with a truly multilingual publishing history.

Teaching through a minimal computing lens, moreover, greatly benefits from projects that exhort students to think both globally and locally by recognizing the technological affordances they have access to (as well as why and how) and by confronting the limitations and constraints

that work against them, whether in hardware, software, education, network capacity, power, or indeed self-imposed limitations for pedagogical purposes. The nature of the cross-border collaboration between students is online and virtual, given their geographical separation. They attend virtual lectures and collaborate online via messaging and code sharing platforms, with the support of the instructors. This kind of engagement is often referred to as "Virtual Exchange" (Bassani and Buchem, 2019; O'Dowd, 2018) or "Collaborative Online International Learning" (COIL) (Guth, 2013).

Our continued challenge is uncovering how our work teaching minimal computing can effectively advance a more open and global digital humanities. We aim at moving beyond the limits of the course itself, by upholding approaches to pedagogy and digital humanities research that work towards what we claim should be a core tenet for the global digital humanities community: technology owned by no one and used and contributed to by all.

Bibliography

Arancio, J. C. (2021). *Opening up the tools for doing science: The case of the global open science hardware movement.* International Journal of Engineering, Social Justice and Peace, 8(2), pp. 1-27. https://ojs.library.queensu.ca/index.php/IJESJP/article/view/13997/9834.

Bassani, P. S. and Buchem, I. (2019). Virtual Exchanges in Higher Education: Developing Intercultural Skills of Students Across Borders Through Online Collaboration. *Revista Interuniversitaria de Investigación en Tecnología Educativa,* 6. https://doi.org/10.6018/riite.377771.

Guth, S. (2013). *The COIL Institute for Globally Networked Learning in the Humanities*. SUNY - COIL Center-SUNY Research Foundation. http://coil.suny.edu/sites/default/files/case_study_report.pdf.

Minimal Computing Working Group. (2021). *About. What is Minimal Computing?* https://go-dh.github.io/mincomp/about.

O'Dowd, R. (2018). From Telecollaboration to Virtual Exchange: State-of-the-art and the Role of UNICollaboration in Moving Forward. *Journal of Virtual eExchange*, 1: 1–23. https://doi.org/10.14705/rpnet.2018.jve.1.

Sahle, P. (2016). What is a Scholarly Digital Edition? In Pierazzo, E. and Driscoll, M. *Digital Scholarly Editing: Theories and Practices*, pp. 19–40. Cambridge, UK: Open Book Publishers. https://jstor.org/stable/j.ctt1fzhh6v.6.

UNESCO, **Leeds-Hurwitz, W. and Stenou, K.** (2013). *Intercultural Competences: Conceptual and Operational Framework.* https://unesdoc.unesco.org/ark:/48223/pf0000219768.

DeXTER: A post-authentic approach to heritage visualisation

Viola, Lorella
lorella.viola@uni.lu
University of Luxembourg, Luxembourg

INTRODUCTION

Cultural heritage institutions and academics are resorting more and more to visual representations of cultural heritage material as a way to enhance access to collections for users' appreciation and research purposes (Windhager et al., 2019). However, scholars have pointed out (Drucker, 2011; 2013; 2014; 2020; Windhager et al., 2019) how a critical approach to visualisation is still largely missing and how on the whole, user-interface (UI) design still shows a functional and task-driven approach, oriented towards satisfying a need for information rather than towards eliciting curiosity, engagement and reflection. With this presentation, I aim to contribute to the urgent need for the establishment of a critical data and visualisation literacy in current task-driven approaches to interface design which continue to see the user as a consumer and to operate within a problem solver model. Drawing on critical posthumanities (Braidotti, 2017; 2019), I here propose a critical framework to digital cultural heritage and digital cultural heritage visualisation. With this approach —which I labelled "post-authentic framework" (Viola, 2021) —I want to initiate a discussion and critique of the fetishization of empriricism and technical objectivity in digital knowledge creation.
To exemplify how the post-authentic framework works in practice, I present the design choices for developing the tool DeXTER – DeepteXTminER, an interactive visualisation app to explore enriched cultural heritage material. The discussion will revolve around the challenges facing product design, with specific reference to visualising the ambiguities and uncertainties of network analysis (NA) and sentiment analysis (SA). DeXTER is currently loaded with Chroniclitaly 3.0 (Viola and Fiscarelli, 2021a), a digital heritage collection of Italian American newspapers published in the USA by Italian immigrants between 1897 and 1936.

POST-AUTHENTIC FRAMEWORK

My argument for a post-authentic framework to digital cultural heritage builds upon recent digital heritage positions (see for instance Cameron, 2021; Goriunova, 2019; Jones et al., 2018) and extends on posthuman critical theory which understands the matter as an extremely complex assemblage of "forces, entities, and encounters" (Braidotti, 2017, 16). A post-authentic framework to digital cultural heritage understands curatorial interventions as cyclic processes that shape and are shaped by humans, entities, and processes connected to each other according to the various forms of power embedded in computational processes and beyond. As these processes are never neutral, their implementation requires constant critical supervision. Whilst exploiting the new opportunities offered by computational technologies, a post-authentic framework rejects an uncritical adoption of digital methods and it contests the main discourse that still presents such techniques and outputs as exact, final, objective and true.

NETWORK ANALYSIS – USEFUL BUT PROBLEMATIC

NA is a method that models pairwise relations (i.e., edges) between objects (i.e., nodes) (Biggs, Lloyd, and Wilson, 1986). When applied to cultural heritage material, NA is typically used to investigate how referential entities (i.e., people, organisations, places) are connected with each other. However, the guiding assumption that modelling how entities relate to each other provides adequate explanations of social phenomena conceals the fact that the results are highly dependent on which entities are modelled and more importantly, which ones are not. NA therefore tends to present as accurate conclusions that may be based on over-represented actors or conversely, on underrepresented categories. When NA is applied to cultural heritage material, this issue is particularly significant as the resulting analysis can impact on the provided interpretation of the past.

Another critical issue of NA is related to the intrinsic nature of the technique itself. Nodes are understood as discrete objects, i.e., completely independent from each other, meaning that they are modelled to remain stable regardless of the relationship between them. This offers a rather artificial view of the analysed phenomena as in reality, actors are always affected by mutual relations (Drucker, 2020, 180). Although extremely significant, this issue is often obfuscated by attractive visualisations which present as accurate what in reality is only approximated information (ibid., 87).

DEXTER - INTERFACE DESIGN AND NETWORK ANALYSIS VISUALISATION

DeXTER introduces a counter narrative in the main discourse —public and academic —by documenting, explaining and motivating all the data selections and interventions in the project's GitHub repository (Viola and Fiscarelli, 2021b). In this way, the management of data is acknowledged as problematic, an act of constant interpretation and manipulation which transforms, selects, aggregates, and ultimately creates data. In the case of NA, DeXTER openly acknowledges that the displayed entities are not ALL the entities in the collection but in fact a representative, yet small selection. This choice –as explained in the app's documentation –had the advantage to provide a less crowded network graph; on the other hand, however, it resulted in a loss of representation of the less occurring entities.

INFORMATION RETRIEVAL

DeXTER's interface allows users to carry out efficient information retrieval. As said earlier, NA visualisations are typically static, and nodes and edges remain unchanged no matter their relation. In DeXTER it is always possible to explore the changing relationships between entities over time or at specific intervals; this is done by swiping the time bar on the top left of the interface which reflects the collection's timespan, i.e., 1897-1936 (Figure 1). This allows users to obtain a more realistic representation of how the modelled entities were mentioned by migrants over time or at specific points in history. Users can also select/ deselect specific titles and observe how entities of interest were mentioned by migrants in titles of different political orientation and geographical location.

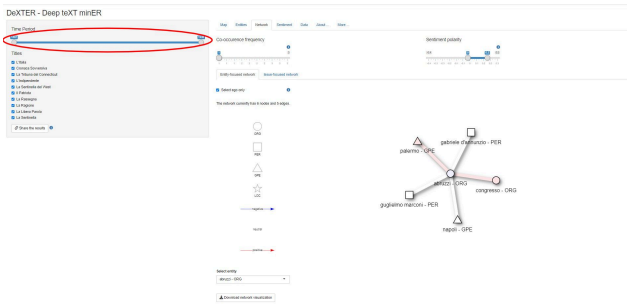


Figure 1:
DeXTER default landing interface for NA. The red oval highlights the time bar

DATA MULTI-DIMENSIONALITY

DeXTER exposes the data multi-dimensional complexity by allowing users to explore three types of networks: two entity-focused graphs and one issue-focused network. In the entity-focused graph (i.e., ego-network), users start from a selected entity of their choice (i.e., the ego) to explore the net of entities (i.e., the alters) most frequently mentioned in the same sentence as the selected entity. The network also displays the number of times entities were mentioned together, the titles in which they were mentioned and the prevailing emotional attitude of the sentence in which they were mentioned. If ego-network is not selected, the graph displays also the relations among the alters. Figures 2 and 3 show the different entity-focused displays of the same entity/ego when selecting/deselecting the ego-network option. Finally, it is possible to obtain additional points of view by starting from a specific issue (i.e., issue-focused network) so as to research which actors were mostly mentioned on a specific day (or days) by a specific newspaper or several (Figure 4).

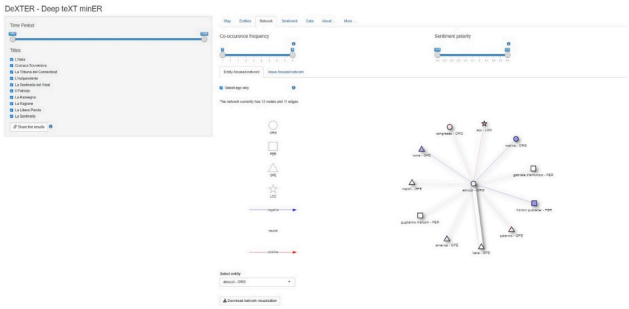


Figure 2:
DeXTER entity-focused NA - egocentric network

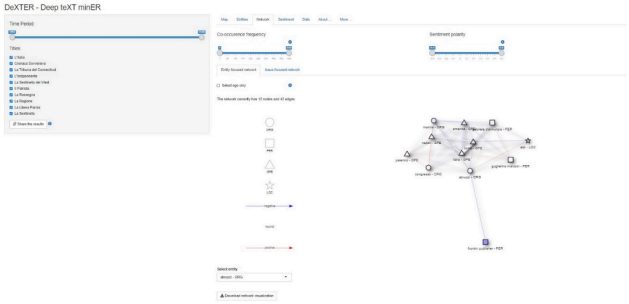


Figure 3:
DeXTER entity-focused NA

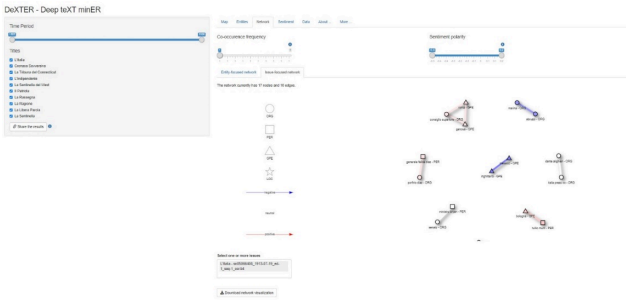


Figure 4:
DeXTER issue-focused NA

SENTIMENT ANALYSIS

SA aims to identify the prevailing emotional attitude in a given text, e.g., positive, negative, or neutral. However, these labels are typically presented as if they were unambiguous categories universally accepted. As a way to acknowledge the ambiguities behind a 'sentiment score', we implemented a fluid visualisation through colour gradients which go from a darker shade of blue for the lowest score (i.e., negative) to a darker shade of red for the highest score (i.e., positive). We also chose pastel shades as opposed to solid shades as an open way to acknowledge SA as problematic and to promote a visualisation that would not be presented as precise and accurate. The description of how the sentiment categories have been identified, how the classification has been conducted, what the scores mean and how they have been rendered in the visualisation, and how the results have been aggregated is been thoroughly documented in the dedicated OA GitHub repository (Viola and Fiscarelli, 2021b).

DATA ACCESSIBILITY

Finally, the data behind the interface is accessible and downloadable through the tab 'Data' which remains always active from anywhere in the interface. Users can choose to either download the entire data-set or the data-set that reflects their specific selections operated through the available filters (e.g., title, time interval, frequency, entity), which also remain visible at all times. The intention is to emphasise the continuous making and re-making of data and to encode this process of forming, arranging and interpreting data within the interface itself. Moreover, in this way, DeXTER also advocates transparency, traceability, and accountability. Indeed, a post-authentic framework to digital cultural heritage primarily acknowledges the collective responsibility of building a source of knowledge for current and future generations and frames it as honest and accountable, unfinished and receptive to alternatives.

Bibliography

Biggs, Norman, E. Keith Lloyd, and Robin J. Wilson. (1986). *Graph Theory*, 1736-1936. Oxford [Oxfordshire] ; New York: Clarendon Press.

Braidotti. (2017). "Posthuman Critical Theory." *Journal of Posthuman Studies* 1 (1): 9. https://doi.org/10.5325/jpoststud.1.1.0009.

Braidotti, Rosi. (2019). "A Theoretical Framework for the Critical Posthumanities." Theory, *Culture & Society* 36 (6): 31–61. https://doi.org/10.1177/0263276418771486.

Cameron, Fiona. (2021). *The Future of Digital Data, Heritage and Curation in a More-than Human World*. Abingdon, Oxon; New York, NY: Routledge.

Drucker, Johanna. (2011). "Humanities Approaches to Graphical Display." *Digital Humanities Quarterly* 5 (1). http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html.

———. (2013). "Performative Materiality and Theoretical Approaches to Interface." DHQ: *Digital Humanities Quarterly* 7 (1).

———. (2014). "Visualizing Temporality: Modelling Time from the Textual Record." Queen Mary University of London, March 25. http://www.qmul.ac.uk/events/items/2014/119740.html.

———. (2020). *Visualization and Interpretation: Humanistic Approaches to Display*. Cambridge, Massachusetts: The MIT Press.

Goriunova, Olga. (2019). "The Digital Subject: People as Data as Persons." *Theory, Culture & Society* 36 (6): 125–45. https://doi.org/10.1177/0263276419840409.

Jones, Siân, Stuart Jeffrey, Mhairi Maxwell, Alex Hale, and Cara Jones. (2018). "3D Heritage Visualisation and the Negotiation of Authenticity: The ACCORD Project." *International Journal of Heritage Studies* 24 (4): 333–53. https://doi.org/10.1080/13527258.2017.1378905.

Viola, Lorella. (2021). "Keynote: The Importance of Being Digital." In *HistoInformatics2021 Workshop* - 6th International Workshop on Computational History. Co-Located with JCDL2021. University of Illinois. https://sites.google.com/view/histoinformatics2021workshop/home?authuser=0.

Viola, Lorella, and Antonio Maria Fiscarelli. (2021a). *ChroniclItaly 3.0. A deep-learning, contextually enriched digital heritage collection of Italian immigrant newspapers published in the USA, 1898-1936*. Zenodo. https://doi.org/10.5281/ZENODO.4596345.

Viola, Lorella, and Antonio M. Fiscarelli. (2021b). *DeXTER (DeepTextMiner): A Deep Learning, Critical Workflow to Contextually Enrich Digital Collections and Visualise Them* (version v1.0.0). Zenodo. https://doi.org/10.5281/ZENODO.4633404.

Windhager, Florian, Paolo Federico, Gunther Schreder, Katrin Glinka, Marian Dork, Silvia Miksch, and Eva Mayr. (2019). "Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges." *IEEE Transactions on Visualization and Computer Graphics* 25 (6): 2311–30. https://doi.org/10.1109/TVCG.2018.2830759.

Textbooks and space – a tale of two dimensions. A GIS analysis of cultural content of language textbooks.

Wacławik, Paulina
paulina.h.waclawik@gmail.com
University of Warsaw, Poland

Culture has been variously conceptualised by the researchers be it as an onion (Hofstede *et al.*, 2010), an iceberg (Hall, Edward, T., 1976), a (Moran, Patrick, R., 2006), a set of binary categories (Hofstede, 2001), or a spectrum of language and culture (Risager, 2007). Thus, to study its foundations, we need to address values, and to do so, we have to resort to investigating cultural artefacts. One particularly telling example of an artifact of culture are language textbooks (Gray, 2010; Gray, 2000)

In my study, I adjust the method of textbook analysis proposed by, e.g., Karen Risager (2018) and Michael Hollenback's adaptation of Byram's Model of Intercultural Citizenship (MICC) (Hollenback; Hollenback, 2016; Byram, 1997) to the language of spatial analysis in Geographic Information Systems, GIS in short. As a case study, I use academic English textbook _ The World We Live In (2013) published for Japanese students in Japan by Japanese publishing house (reflecting the idea of taking into account the three parties involved in the creation of the textbooks: authors, teachers, and students). The textbooks are investigated from the perspective of space (understood as the space of the textbooks, which is the arrangement of the content on the pages of the book) and space depicted in the books (i.e., geographical space). The two levels of space will translate to two levels of analysis: the construction of the book in terms of e.g., its themes, types of tasks, and cognitive difficulty;

the role of words with spatial reference (e.gl, placenames, landmarks, people)

which will finally be combined and cross referenced, though heeding the differences in the types of the data acquired (c.f. similar approach to double levels of analysis of textual data in the project on mapping place names in

historical court books by Polish Academy of Science https://atlasfontium.pl/?page_id=2303).

First, the books were scanned and entered into the GIS system. Then, the information in them was coded and marked using GIS, where it is stored in dedicated files, here: hapefiles (mostly polygon and point ones). The layers with shapefiles cover particular content type, for example:

- placenames, including names of the countries, regions, cities (coded according to GeoNames ontology (www.geonames.org), as well as words with other spatial reference, like landmarks or established products;
- thematic analysis of the units or sections of the textbooks (according to Common European Framework of Reference (Council of Europe, Language Policy Unit));
- national and transnational institutions;
- cognitive level of tasks (Bloom's (Anderson and Krathwohl, 2001) revised taxonomy);
- or language used to describe culture (by Patrick Moran (2006)).

Finally, the collected data are subject to spatial analysis using such tools as spatial join, kriging, buffers and distance analysis to investigate, spatial relationships of particular types of content of the textbooks.

The elaborated approach allows to pose new questions and collapse previously separate, often incomparable, discourses into one analysis, as proposed in *Representations of the World in Language Textbooks: Languages for Intercultural Communication and Education (Risager, 2018)* Mapping the content of the textbooks both onto each other and looking for the correlation between the linguistic component and geographical space presented in the books allows to offer a new voice in the discussion on the place of culture in language teaching, often labelled as opposition between English as a Foreign Language and English as a Lingua Franca (EFL/ELF).

We also sought an answer to the question on the impact of Globalization on English teaching by searching for the cultures of interest (understood as places mentioned in the textbooks) and assumed target culture(s) (the English-speaking countries) along with the context they appear in. It is operationalized as: mentions of places and cognitive levels, types of tasks and language used to describe them.

Bibliography

References

Anderson, L. W. and Krathwohl, D. R. (2001). A taxonomy for learning, teaching, and assessing. A revision of Bloom's taxonomy of educational objectives. New York: Longman.

Block, D., Gray, J. and Holborow, M. (2012). Neoliberalism and applied linguistics. London: Routledge.

Byram, M. (1997). Teaching and Assesing Itercultural Communicative Competence. Clevedon: Multilingual Matters.

Gray, J. (2000). The ELT coursebook as cultural artefact: how teachers cenesor and adapt. ELT Journal. 54(3): 274–283.

Gray, J. (2010). The Construction of English. Culture, Consumerism and Promotion in the ELT Global Coursebook. Houndsmills Basingstoke Hampshire, New York NY: Palgrave Macmillan.

Hall, Edward, T. (1976). Beyond Culture. Anchor Books.

Hofstede, G. (2001). Culture's consequences_ comparing values, behaviors, institutions, and organizations across nations. Sage.

Hofstede, G., Hofstede, G. J. and Minkov, M. (2010). Cultures and organizations. Software of the mind : intercultural cooperation and its importance for survival. New York: McGraw-Hill.

Hollenback, D. M. (ed.). Incorporating Intercultural Communication into EFL Education in Japan Utilizing Task-Based Learning Course Design.

Hollenback, D. M. A Critical Look at Culture in EFL Textbooks in Japan, Transformation in language development, Nagoya, Japan, 2016.

Ogasawara, S., Hiroe, A. and Cutrone, P. (2013). The World We Live In. Tokyo: EINOUSHA.

Risager, K. (2007). Language and culture pedagogy. From a national to a transnational paradigm. Clevedon, Buffalo: Multilingual Matters.

Risager, K. (2018). Representations of the World in Language Textbooks. Languages for Intercultural Communication and Education). Bristol UK: Multilingual Matters.

Mining an Interpolated Commentary for Linguistic Markup

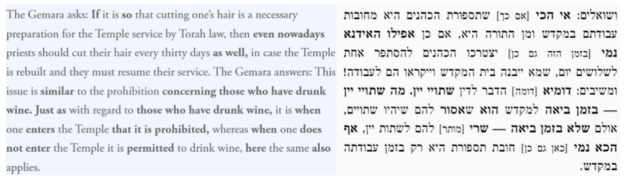
Waxman, Joshua
joshua.waxman@yu.edu
Stern College for Women, Yeshiva University, United States of America

Introduction

The Babylonian Talmud is a multi-volume legal, ethical, narrative, and religious work containing scholastic discourse spanning several centuries and two countries - Iraq and Israel. It is written in a mixture of Biblical Hebrew, Biblical Aramaic, Middle Hebrew and Babylonian Aramaic, all Semitic languages written in Hebrew characters. The Talmud is the focus of religious and academic study, and there have been several Talmud-related digital humanities projects in recent years, e.g. (Waxman, 2021a; Satlow and Sperling, forthcoming; Zhitomirsky-Geffet and Prebor, 2019; Jutan and Regenbaum, 2019).

However, the Talmudic text is difficult for both humans and machines to parse, for several reasons. The Hebrew alphabet is consonantal; vowels are represented by optional diacritics, which are missing in the Talmud. Hebrew words are highly inflected, and it's sometimes difficult to separate stem from morphology. בצל could mean "in shade", "in the shade", "onion", or "onion of". Each word is thus heavily ambiguous, and each sentence exponentially so. The Talmudic corpus contains more than 1.8 million words with a unique linguistic profile. Its multiple languages, each with unique vocabulary and word forms, introduces further ambiguity. It's unpunctuated, so sentence and phrase boundaries are uncertain, introducing ambiguity into dependency or constituent parses and NLP tasks such as NER and relation extraction. The Talmud frequently code-switches between its four Semitic languages, increasing ambiguity / incomprehensibility. Our prior Digital Humanities work (Waxman, 2021a) compensated by focusing primarily on an aligned English translation and projected to the Hebrew side, but there are good reasons to focus on the original text rather than translation. How can we build richly-tagged Talmudic corpora and associated linguistic tools?

Sefaria, an open-source Jewish library, includes a digital edition of the Talmud. They segmented the text of the Talmud into paragraphs, and aligned many rabbinic commentaries to these paragraphs. Among these commentaries are Steinsaltz's English commentary and his Modern Hebrew Commentary. Each commentary is "interpolated", meaning that there is the literal Talmudic text (bolded), which is extremely concise in style, as well as elaborative gloss text (unbolded), turning brief cryptic statements into well developed ideas and sentences. One such interpolated paragraph, drawn from Taanit 17a, appears in Figure 1.


The Gemara asks: **If** it is **so** that cutting one's hair is a necessary preparation for the Temple service by Torah law, then **even nowadays** priests should cut their hair every thirty days **as well,** in case the Temple is rebuilt and they must resume their service. The Gemara answers: This issue is **similar** to the prohibition **concerning those who have drunk wine. Just as** with regard to **those who have drunk wine,** it is **when** one **enters** the Temple **that it is prohibited,** whereas **when** one **does not enter** the Temple it is **permitted** to drink wine, **here** the same **also** applies.

ושואלים: **אי הכי** [אם כך] שתספורת הכהנים היא מחובות עבודתם במקדש ומן התורה היא, **אם כן אפילו האידנא נמי** [בזמן הזה גם כן] יצטרכו הכהנים להסתפר אחת לשלושים יום, שמא ייבנה בית המקדש וייקראו הם לעבודה! ומשיבים: **דומיא** [דוגמה] הדבר לדין **שתויי יין. מה שתויי יין** — **בזמן ביאה** למקדש **הוא שאסור** להם שיהיו שתויים, אולם **שלא בזמן ביאה** — **שרי** [מותר] להם לשתות יין, **אף הכא נמי** [כאן גם כן] חובת תספורת היא רק בזמן עבודתה במקדש.

Figure 1: Interpolated Steinsaltz Talmud commentary, Taanit 17a

The Hebrew interpolated commentary doesn't precisely match Talmud's original mix of Hebrew and Aramaic. Sometimes, it draws from variant textual sources, so words and phrases are missing. Since Steinsaltz's goal is to produce flowing Hebrew sentences, in Hebrew / Aramaic hybrid words and cognates, he'll sometimes replace Aramaic prefixes with their Hebrew equivalent. Thus the Talmud's דאסור became שאסור in Figure 1. Some abbreviations and acronyms are expanded in the commentary. Still, we realized that we could word-align the Talmud with commentary and project linguistic features to create a rich Talmudic text.

Approach

We first tokenized the Hebrew commentary, separating off punctuation and gloss, and word-aligned it with the original Talmudic text. Match-reward was based on overlap of character bigrams of the source and target words.

We projected punctuation from the Hebrew commentary to the Talmudic text, recovering punctuation embedded within the gloss and applying a heuristic by which stronger punctuation replaced weaker punctuation (e.g. period for comma). By analogy, if the bolded English words in Figure 1 were Talmudic text, we would produce the sentence: **If so, even nowadays as well.** and matching pairs of quotation marks surrounding literal text are selected for projection as well. We segmented the now-punctuated sentences using the projected periods, question marks, exclamation marks, and interrobangs (Waxman, 2022).

Steinsaltz deals with arcane Middle Hebrew words / phrases by bolding the original Hebrew and following it with unbolded Modern Hebrew translation, in parentheses. For Aramaic words / phrases, he gives the bolded original and provides the Modern Hebrew in brackets, with literal translation in small text (Figure 1) and glossed prefixes and phrases in regular-sized text. He also employs parentheses to source Biblical verses.

We next developed a rough heuristic to mark Talmudic words for language. Words default to Hebrew. When encountering open parentheses, if the following word is a Hebrew Biblical book, the words until the open quote are marked as Biblical Hebrew. If a partly Aramaic book, the quoted phrase is marked as Biblical Aramaic. When encountering open brackets, the preceding phrase until

punctuation is marked as Aramaic. We trained an RNN-CRF model on this corpus to perform language identification for new texts (Waxman, 2021b).

אי הכי	אם כך
אפילו האידנא נמי	בזמן הזה גם כן
דומה	דומיא
דשתויי	שתויי
יין	יין

Figure 2: Hebrew translation corpus, Taanit 17a

Our new work builds on this. We produce an aligned translation corpus. Words are generally word-aligned, but we align full parenthesized / bracketed phrases to align to words / phrases (see Figure 2). For instance, the Aramaic phrase אי הכי, meaning "if so", is paired with the equivalent Hebrew phrase אם כך, and the individual Aramaic word יין, "wine" is paired with its equivalent Hebrew יין. Some errors are present, for instance the prefix ד, "of" in דשתויי, wasn't present in the Hebrew commentary. This translation corpus is input for our POS tagging. We employ Dicta's API (Shmidman et al., 2020) to obtain ranked POS and morphological analyses for the original Talmudic text, selecting the top Aramaic choice when we know the word is Aramaic. We also POS-tag the translated Hebrew side. Figure 3 shows an Aramaic phrase, where אין means "yes/ indeed", the same as Hebrew כן. Where there is overlap of POS options, we've eliminated some ambiguity and rerank the POS selections.

Aramaic		Hebrew	
אין	EXISTENTIAL	כן	ADVERB
ומצאתה	VERB	ומצאתה	VERB
דקרא	NOUN	שבכתוב	NOUN
דאתא	VERB	שבא	VERB
לידיה	NOUN	לידו	PREPOSITION
משמע	PARTICIPLE	משמע	PARTICIPLE

Figure 3: Hebrew and Aramaic POS-tagged with differences

Finally, we perform discourse classification of our segmented sentences, to classify sentences as statements, questions, or answers. Features include sentence-initial and final words / phrases, punctuation, and commentary's framing words such as ושואלים / ומשיבים ("they ask / answer" - see the first Hebrew word in Figure 1). Future work involves tweaking heuristics to handle more edge cases and improve accuracy, particularly in quoted text, and developing semantic grammars to discover named entities / relations from our punctuated sentences.

Results

Our punctuation results are generally in the 90-95% range for precision and recall (see Table 1). Quotes have 95.6% recall and 93.1% precision.

		generated punctuation									
		,	;	:	—	.	?	?!	!	null	recall
gold punctuation	,	4884	9	39	20	90	7	0	3	206	92.89%
	;	3	1	0	0	3	0	0	0	1	12.50%
	:	51	1	3692	8	66	2	0	2	219	91.36%
	—	27	0	12	1816	12	5	0	0	193	87.94%
	.	101	5	39	3	4817	15	3	29	64	94.90%
	?	7	0	0	1	8	793	2	1	15	95.89%
	?!	0	0	0	0	0	2	119	0	1	97.54%
	!	4	0	0	0	2	3	0	341	9	94.99%

Table 1: Confusion matrix for punctuation

Table 2 shows accuracy for our language identification process, using a CRF model. Using an RNN-CRF model we achieve 95% accuracy (93% on non-punctuation). Our POS tagging / discourse analysis are harder to assess on a large scale, but we'll publish the resulting corpora.

	precision	recall	f1-score	support
Aramaic	0.95	0.93	0.94	114890
Hebrew	0.95	0.96	0.95	240711
BiblicalHebrew	0.75	0.71	0.73	14481
BiblicalAramaic	0.71	0.56	0.63	244
_	1.00	1.00	1.00	118958
accuracy			0.95	489284
macro avg	0.87	0.83	0.85	489284
weighted avg	0.95	0.95	0.95	489284

Table 2: CRF model, 200 epochs

Bibliography

Jutan, D. and Regenbaum, S. (2019). The Future of the Talmud: A Digital Humanities Case Study. Georgia Tech. https://dilac.iac.gatech.edu/node/66

Satlow M. and Sperling, M. (forthcoming). The Rabbinic Network. In *AJS Review*.

Shmidman, A., Shmidman, S., Koppel, M., and Goldberg, Y.(2020). Nakdan: Professional Hebrew Diacritizer. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.* pp. 197-203:

Waxman, J. (2021a). A graph database of scholastic relationships in the Babylonian Talmud. In Quené et al. (eds), *Digital Scholarship in the Humanities.* Oxford University Press, pp. ii277–ii289.

Waxman, J. (2021b). Projecting and Detecting Code Switching in the Babylonian Talmud. Presentation at ACH 2021.

Waxman, J. (2022). Projecting Punctuation from an Interpolated Translation and Commentary. In Chesner et al (eds), *Jewish Studies in the Digital Age*, *Studies in Digital History and Hermeneutics*. De Gruyter.

Zhitomirsky-Geffet, M. and Prebor, G. (2019). SageBook: Toward a cross-generational social network for the Jewish sages' prosopography. In *Digital Scholarship in the Humanities*. Oxford University Press, pp. 676–695.

The Many Voices of Du Ying: Revisiting the Disputed Writings of Lu Xun and Zhou Zuoren

Xie, Xin
rwxiexin@shnu.edu.cn
Shanghai Normal University, Shanghai, China

Wang, Haining
hw56@indiana.edu
Indiana University Bloomington, Bloomington, Indiana, USA

Riddell, Allen
riddella@indiana.edu
Indiana University Bloomington, Bloomington, Indiana, USA

Summary

We revisit essays by Lu Xun (鲁迅) and Zhou Zuoren (周作人) that were initially published pseudonymously. The authorship of several of these essays is disputed. Through a quantitative analysis of the author's writing styles, we find evidence supporting the following:

1. Two pseudonyms, Du Ying (独应) and Du (独), are used by both authors,
2. The *On the Difference Between the Russian Revolution and Nihilism* (论俄国革命与虚无主义之别) and seven other essays are likely written by Lu Xun,
3. The *People of Yue, Forget Not Your Ancestors' Instructions* (尔越人毋忘先民之训) and two additional essays are likely written by Zhou Zuoren, and
4. the authors may have written *The Strings of Melancholy* (哀弦篇) and another two essays collaboratively.

Backgrounds

Lu Xun (1881–1936) is perhaps the best-known writer in modern China. His essays feature in textbooks used in secondary and tertiary education across East Asia. As a short story writer, essayist, and critic, Lu Xun's work is known for its incisive social commentary as well as its use of irony and satire.

Although several compilations of Lu Xun's writing exist, completing a comprehensive collection of his works is a challenge because so many of his works were written under a pseudonym. Over the course of his writing career, Lu Xun employed at least 150 pen names (Xu and Hong, 1988). Nobody—including Lu Xun himself—kept precise records of his publications. Lu Xun's frequent use of pseudonyms primarily served to help him evade censorship from the government of the Republic of China. But he also used a pseudonym for other reasons, such as to express his dissatisfaction with society. Finally, some writings were forged for effect or for other reasons, e.g., *Literary Chatters* (艺文杂话) (Ding and Ding, 1989). Further complicating efforts to assemble a comprehensive collection of his works is the fact that the authorship of several of his essays is actively disputed: four articles signed with the pen name Du Ying are also claimed by Zhuo Zuoren (1885–1967). Zhou is a younger brother of Lu Xun. Both studied in Japan in their youth and later led the New Culture Movement advocating writing using vernacular Chinese.

Peng and Ma (1981) assign four essays signed by Du Ying to Lu Xun, but none of them can be found in the *Complete Collection of Lu Xun* (鲁迅全集), a compilation published in the same year (Lu Xun, 1981). Zhong Shuhe includes these essays in the *Complete Prose Collection of Zhou Zuoren* (周作人散文全集) (Zhou, 2009). Interestingly, the essays are not found in the collection that Zhou edited himself (the *Self-edited Collection of Zhou Zuoren*, 周作人自编文集 (Zhou, 2002)).

Other opinions about the identity of Du Ying exist. Chen (1980) concludes the pseudonym can be used by both authors. Chen grounds his conclusion by studying the development of each writer's thinking over time. Collaboration between the two brothers is another possibility (Meng, 2017).

Research Question

Who wrote under the pseudonym Du Ying? Is it Lu Xun or Zhou Zuoren—or has the pseudonym been used in a collaboratively written work? To begin to tackle this problem, we frame this research question as one of authorship attribution (Juola, 2006). Authorship attribution

techniques infer the likely author of an unsigned or disputed text based on quantitative analysis of lexical and syntactic features found in the text. Standard authorship attribution assumes:

• The candidate set is known, e.g., two possible authors in our case.

• The unsigned text is singularly authored, i.e., no collaboration or substantial editorial moderation is present.

Authorship attribution has been applied in numerous cases. Notable examples include the Federalist papers (Mosteller and Wallace, 1963), disputed plays between Shakespeare and Fletcher (Matthews and Merriam, 1993), and the rabbinic responsa *Torah Lishmah* (Koppel et al., 2007), and disputed Latin Visons the *Visio ad Guibertum missa* and *Visio de Sancto Martino* (Kestemont et al., 2015) to name a few.

Corpus

We begin by organizing a corpus suited to the problem. Instead of only looking at the four disputed essays, we retrieve all texts published under the pen names Du Ying and Du. These essays were published at roughly the same time and in similar publication venues. We are, in effect, expanding the range of texts we intend to consider using authorship attribution analysis. This is appropriate as we are dealing with texts whose authorship is at least somewhat uncertain. Twenty-one essays in total are recovered and comprise the "test set," the essays whose authorship we intend to predict. Table 1 lists the texts.

Author	Purpose	Title	Topic	Length	Chunks
Lu Xun	Train	Lessons from the History of Science (科学史教篇)	history & science	7,032	7
	Train	On the Aberrant Development of Culture (文化偏至论)	culture & politics	8,556	7
	Validation	On Radium (说镭)	science	3,094	4
	Validation	On the Power of Mara Poetry (摩罗诗力说)	literary & politics	23,779	22
Zhou Zuoren	Train	Preface to Midst the Wild Carpathians (《匈奴奇士录》序)	literary	627	1
	Train	Preface to Charcoal Drawing (《炭画》序)	literary	258	1
	Train	Preface to The Lost History of Red Star (《红星佚史》序)	literary & history	1,145	1
	Train	Preface to The Yellow Rose (《黄薔薇》序)	literary	786	1
	Train	A Brief Discussion on Fairy Tales (童话略论)	literary & history	3,093	4
	Train	A Study on Fairy Tales (童话研究)	literary & history	5,083	6
	Validation	Preface to Qiucao Garden Diary (《秋草园日记》序)	history	178	0.7
	Validation	An Addendum to Yisi Diary (乙巳日记附记一则)	culture & history	63	0.3
	Validation	A Glimpse of Jiangnan Examinees (江南考先生之一斑)	history	147	0.5
	Validation	Plight and Broil in a Steamboat (汽船之窘况及苦热)	history	144	0.5
Du Ying	Test	The Issue of Women's Suffrage (妇女选举权问题)	politics	1,068	2
	Test	The Issue of Women's Suffrage, Continued (妇女选举权问题续)	politics	1,494	3
	Test	George Eliot (乔治爱里阿德)	literary & history	618	1
	Test	Prisoners of Siberia (西伯利亚之囚)	literary & history	893	1
	Test	Eliza Orzeszko (爱理萨阿什斯珂)	literary & history	322	1
	Test	Stepniak (斯谛勃骂克)	literary & history	459	1
	Test	Petofi (斐彖飞)	literary & history	489	1
	Test	The Power of Writing (文章之力)	literary	720	1
	Test	An Extraordinary Stratagem to Prevent Illicit Sex (防淫奇策)	culture & politics	872	2
	Test	The Chinese Patriotism (中国人之爱国)	history & politics	857	1
	Test	Reaction to the Prison Book Sold at the Store (见店头监狱书所感)	history & politics	925	1
	Test	On the Difference Between the Russian Revolution and Nihilism (论俄国革命与虚无主义之别)	history & politics	2,602	3

	Test	Translator's Note to Silence (《寂漠》译记)	literary & history	348	1
	Test	Translator's Note to At a Country House (《庄中》译记)	literary	192	1
	Test	The Strings of Melancholy (哀弦篇)	literary & history	13,112	21
	Test	Looking at the Land of Yue (望越篇)	history & politics	741	1
	Test	Looking at the Country of China (望华国篇)	history & politics	749	1
Du	Test	People of Yue, Forget Not Your Ancestors' Instructions (尔越人毋忘先民之训)	history & politics	317	1
	Test	Where Has the Character of the Republic Gone? (民国之征何在)	history & politics	354	1
	Test	The Ordinary Folks' Responsibility (庸众之责任)	history & politics	334	1

Table 1: Summary of the corpus containing works of Lu Xun, Zhou Zuoren, "Du Ying", and "Du". Four distinct validation samples of Zhou are combined to make two longer texts.

We then search for texts which we are certain give us information about the brothers' classical writing styles. The texts gathered are written using classical non-fictional prose and were published roughly at the same time relative to the disputed works. With these controls in place, we mitigate concerns about the possible influence on writing style of time (Glover and Hirst,1996), genre (Kestemont et al., 2012), and register (Wang et al., 2021). With the exception of six essays reserved for validation, these texts are used for training. Validation texts are chosen such that they have somewhat different topics than the training samples. Classification accuracy on the validation texts should give us some rough sense of how well the classifier performs. The training and validating texts are from self-edited collections of the authors. Hence their authorship is not in question.

Preprocessing

We carefully remove quotations, dialog, and known text reuse [1]. The longer samples are split into chunks of ca. 800 characters (without breaking paragraphs). Chunking in this fashion does not affect any of the authorship attribution techniques we consider but does let us explore style variation within longer texts. Given that previous research has raised the question of collaborative authorship, this is useful (e.g., some essays published in *Tianyi* in 1907 may have been collaborations (Meng, 2017)). Analysis of chunks may capture clues of collaboration even the chunking procedure is somehow arbitrary. For example, if one isolated chunk within a test set essay has high probability of being written by Lu Xun and the other chunks have high probability of being written by Zhou, this essay may merit further investigation as potentially having been a collaboration.

Model

We use a standard authorship attribution model. It uses function word frequencies as features. The particular set of function words is taken from (Wang et al., 2021). Wang et al. (2021) demonstrated that the function word set was useful in predicting authorship in Ming-Qing fiction written in classical Chinese.

Function word usage is telltale signs of writing style. Function words are mostly unrelated to topic (e.g., "the" and "on" in English; "和" and "况且" in Chinese) and have been successfully applied in extensive authorship attribution studies (Kestemont, 2014). Our function word list has 819 function characters/words in total (262 unigrams, 545 bigrams, ten trigrams, and two quadgrams).

The classification model uses standard logistic regression (ℓ_2 regularization, $\lambda = 1.0$). Logistic regression is chosen for its wide use and interpretability. We use the same preprocessing procedure found in the previous research (Wang et al., 2021): function word counts are normalized by document length in characters and scaled to have unit variance after deducting means. [2]

Decision Rule

For longer prose, a decision is made only when the predicted labels of all chunks agree. If an essay is short, the predictive probability of the author should be no less than 0.75. This decision rule counts as conservative, we think. Readers are welcome to use their thresholds.

Results

The model shows 93% accuracy on the validation set. This gives us confidence that it should provide useful information when making predictions about disputed texts.

The summary of the results is shown in Table 2. Clearly the two pseudonyms are shared between by Lu Xun and Zhou Zuoren. *The Power of Writing* and seven other prose appear to be written by Lu Xun. The *Prisoners of Siberia* and another two essays appear to be written by Zhou Zuoren. We withhold judgement about the rest of our decisions.

Pseudonym	Publication Date	Title	Decision	Note
Du Ying	1907.07.25	The Issue of Women's Suffrage	uncertain	labels disagree
	1907.09.15	The Issue of Women's Suffrage, Continued	uncertain	labels disagree
	1907.09.15	George Eliot	uncertain	low confidence
	1907.09.15	Prisoners of Siberia	Zhou	–
	1907.09.15	Eliza Orzeszko	Zhou	–
	1907.10.30	Stepniak	uncertain	low confidence
	1907.10.30	Petofi	uncertain	low confidence
	1907.10.30	The Power of Writing	Lu Xun	–
	1907.11.30	An Extraordinary Stratagem to Prevent Illicit Sex	Lu Xun	–
	1907.11.30	The Chinese Patriotism	Lu Xun	–
	1907.11.30	Reaction to the Prison Book Sold at the Store	Lu Xun	–
	1907.11.30	On the Difference Between the Russian Revolution and Nihilism	Lu Xu	–
	1908.12.05	Translator's Note to Silence	Zhou	-
	1908.12.05	Translator's Note to At a Country House	uncertain	low confidence
	1908.12.20	The Strings of Melancholy	uncertain	labels disagree
	1912.01.18	Looking at the Land of Yue	Lu Xun	–
	1912.01.22	Looking at the Country of China	Lu Xun	–
Du	1912.02.01	People of Yue, Forget Not Your Ancestors' Instructions	Zhou	–
	1912.02.02	Where Has the Character of the Republic Gone?	Lu Xun	–
	1912.02.16	The Ordinary Folks' Responsibility	uncertain	low confidence

Table 2: Summary of the authorship attribution results.

Discussion

We found predicted chunk labels disagree in *The Strings of Melancholy* and other two essays. For example, the fifth chunk of *The Strings of Melancholy* is predicted to be written by Zhou with 0.85 confidence. But the first and second chunks of the essay have over 0.94 probability of being written by Lu Xun. A similar phenomenon also happens in *The Issue of Women's Suffrage* and *The Issue of Women's Suffrage, Continued*. Evidence like this could be a hint of collaboration. As Meng (2017) suggested, the brothers read and wrote closely in 1907.

The impact of topics on the classifier deserves further investigation. (Some function words may be more (or less) likely to appear in certain topics.) We do, however, think this concern should not be overstated. For example, *People of Yue, Forget Not Your Ancestors' Instructions* concerns political matters and is assigned to Zhou with high probability. Yet Zhou has few training examples that concern politics. We intend to further investigate the possibly interplay between subject matter and authorial style.

Bibliography

Chen, S. (1980). Revisiting the article signed by "Du Ying" in Tianyi newspaper. *Historical Studies of Modern Literature*, **3**: 115–121.

Ding, J. and Ding, Y. (1989). Literary rogue Shi Jixing in the 1930s and 1940s—various deceptions and false

accusations against Lu Xun, Yu Dafu and others. *Jianghuai Forum*, **2**: 92–104.

Glover, A. and Hirst, G. (1996). Detecting stylistic inconsistencies in collaborative writing. In The new writing environment, pp. 147–168. Springer.

Juola, P. (2006). Authorship Attribution. *Foundations and Trends in Information Retrieval*, **1**(3): 233–334.

Kestemont, M., Luyckx, K., Daelemans, M. and Crombez, T. (2012). Cross-genre authorship verification using unmasking. *English Studies*, **93**(3): 340–356.

Kestemont, M. (2014). Function words in authorship attribution. from black magic to theory? *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pp. 59–66.

Kestemont, M., Moens, S., & Deploige, J. (2015). Collaborative authorship in the twelfth century: A stylometric study of Hildegard of Bingen and Guibert of Gembloux. *Digital Scholarship in the Humanities*, **30**(2): 199–224.

Koppel, M., Schler, J. and BonchekDokow, E. (2007). Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*, **8**(6): 1261–1276.

Lu, X. (1981). *Complete Collection of Lu Xun*. Beijing: People's Literature Publishing House.

Matthews, R. and Merriam, T. (1993). Neural computation in stylometry I: An application to the works of Shakespeare and Fletcher. *Literary and Linguistic computing*, **8**(4): 203–209.

Meng, Q. (2017). Reading and writing in the presence of each other: The Zhou brothers in 1907. *Modern Chinese Literature Research Series*, **3**: 106–115.

Mosteller, F. and Wallace, D. L. (1963). Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, **58**(302): 275–309.

Peng, D. and Ma, T. (1981). On four classical essays signed with Du Ying should be written by Lu Xun. *Journal of Liaoning University: Philosophy and Social Science Edition*, **5**: 22–28.

Wang, H., Xie, X. and Riddell, A. (2021). The challenge of vernacular and classical Chinese crossregister authorship attribution. *Proceedings of the Conference on Computational Humanities Research 2021*, pp. 299–309.

Xu, N. and Hong, Q. (1988). *Pen Name Index of Modern Chinese Writers*. Changsha: Hunan Literature and Art Publishing House.

Zhou, Z. (1999). *Selected Collection of Zhou Zuoren*. 2. Beijing: The Masses Press.

Zhou, Z. (2002). *Self-edited Collection of Zhou Zuoren*. Shijiazhuang: Hebei Education Press.

Zhou, Z. (2009). *Complete Prose Collection of Zhou Zuoren*. Guilin: Guangxi Normal University Press.

Notes

1. We drop the dispute essay, *On the Significances of Writing: The Origins of Its Mission and the Faults of Recent Literary Criticism in China*, for now because it contains (unmarked) quotations from other texts. (Zhou, 1999)
2. The implementation relies on *scikit-learn* (v.1.0.1) and *functionwords* (v.0.6) packages on PyPI.

Japanese Old Maps Online for Promoting Digital Humanities

Yano, Keiji
yano@lt.ritsumei.ac.jp
Ritsumeikan University, Japan

Natsume, Muneyuki
natsume@fc.ritsumei.ac.jp
Ritsumeikan University, Japan

Imamura, Satoshi
imamurarumami285@gmail.com
CAD CENTER Co., Ltd.

Kamata, Ryo
ryo@geolonia.com
Geolonia Inc.

1. Introduction

The field of historical GIS has been driving a branch of digital humanities, collaborating with history, literature, cartography, and place name studies. Historical GIS is a fusion of historical geography and GIS and has been developed as a spatial humanities.

A large amount of geospatial information—such as drawings, maps, registers, statistics, and other data in Japan's Early Modern and Modern eras—remains undigitized and unavailable in GIS form, unshared, and unavailable to the public. In order to advance historical GIS, it will be necessary to promote the digitization of geospatial information that is primarily paper-based, convert it in an appropriate manner to GIS, and share it with the public.

The purpose of this paper is to construct 'ARC Map Portal Database' to serve as a portal to Japan's old maps;

2) ‘Japanese Map Warper’; and 3) ‘Japanese Old Maps Online’. This will make it possible to select old maps for analysis from the ARC Map Portal Database, which allows cross-searching of old maps from multiple collections, to share geo-referenced and GIS-enabled maps of old maps with other sites using “Japanese Map Warper”, and finally to import them into “Japanese Old Maps Online”. You will be able to create detailed maps for analysis of various themes by importing them into "Japan Old Maps Online".

2. ARC Map Portal Database

‘ARC Map Portal Database’, a portal for old maps of Japan, currently facilitates searching and browsing of over 5,000 old Japanese maps across numerous possessing institutions (https://www.dh-jac.net/db/maps/search_portal.php)—among them, the Kyoto Institute, Library and Archives; the Museum of Kyoto; the Kyoto City Library of Historical Documents; Ritsumeikan University Library; the British Library; the Sainsbury Institute for the Study of Japanese Arts and Cultures (SISJAC); the University of British Columbia (UBC); the University of California, Berkeley (UCB); Harvard University, and so on.

The maps from these institutions available to the public on the database can be classified into one of three categories: 1) cases where the maps are digitized by the institutions that possess them; 2) cases where the material used has been released on the web, with the possessing institutions themselves pursuing digitization; and 3) cases where the maps are stored at the individual institutions and have not released on the web, but where the institutions have supported releasing photographs of these articles to share for research and educational purposes.

In these cases, when the possessing institutions did not support general release, access was provided to certain users only for research & educational purposes through ID & password authentication.

Old Japanese maps possessed by overseas institutions lack Japanese-language metadata; conversely, those possessed by Japanese institutions lack English-language metadata. Therefore, bilingual metadata for these documents is currently in the process of being revised.

Building this portal has opened up access to a great number of maps useful in geohistorical research and education to everyone.

3. Japanese Map Warper

Adding spatial information to these old maps would allow them to be overlaid and compared with current maps. For this, we create ‘Japanese Map Warper’ web application (https://mapwarper.h-gis.jp/?locale=en). ‘Map Warper’ is an open-source web application for sharing maps and making them available to the public, developed by geospatial information developer and consultant Tim Waters. The application is capable of applying georeference to uploaded maps, searching metadata such as titles and notes, and performing searches within the displayed base map range, as well as narrowing down maps to a certain range of publication years—in other words, making it possible to perform spatiotemporal searches. Map Warper also enables georeferenced maps to be shared in a variety of formats: GeoTIFF, WMS, XYZ tiles, and more.

As of this writing, in late November 2021, 4,465 maps have been uploaded, including old topographic maps; of these, 2,352 have been georeferenced. A large number of people uploading and georeferencing old maps in this manner could greatly increase this platform's usefulness in research and education regarding the landscapes of the past.

4. Japanese Old Maps Online

By using the vast number of old maps available on ‘the ARC Map Portal Database’ on ‘Japanese Map Warper’, we can expect further acceleration of and advancements in historical GIS. However, ‘Japanese Map Warper’ is not capable of conducting detailed analysis such as overlaying multiple old maps and creating vector data. Map-sharing functions indeed allow maps to be used on desktop applications such as ArcGIS and QGIS; however, in order to promote the maps' usage among researchers who do not specialize in GIS—as well as in school, museums, and other educational venues—it is believed that establishing environments where detailed analysis can be conducted conveniently on the web will be required.

Therefore, in this framework (Figure 1), we built ‘Japanese Old Maps Online’ that enables GIS analysis by freely importing geo-referenced old maps of Japan based on ArcGIS Hub. ArcGIS Hub is a feature of ArcGIS Online (a cloud GIS offered by ESRI) that makes it easy to construct open data sites to share data uploaded to ArcGIS Online (https://japanese-old-maps-online-rstgis.hub.arcgis.com/). It is possible to use ArcGIS Online as a bidirectional platform where private citizens, NPOs, and academic institutions can plan the resolution of common issues through participation and sharing. On the data viewing page, data can be exported in a variety of formats, with CSV, KML, Shapefile, and GeoJSON selectable. It is also possible from here to switch to the ArcGIS Online map creation screen directly, create a map using the target data, overlay multiple maps, and (with an ArcGIS paid account) analyse the data with advanced spatial analysis.

5. Conclusion and further challenges

Through this framework, it is now possible even for GIS novices to work with data or old maps they themselves or other users have georeferenced easily, through the Internet —as well as to conduct analysis using a variety of maps and data. Establishing this sort of environment should lead to further advancements in historical GIS research and education. Future issues include making the three systems multilingual and streamlining the transfer of data between them. In addition, we plan to expand the system into applied

research, such as comparative analysis using geo-referenced old maps and automatic extraction of place names.

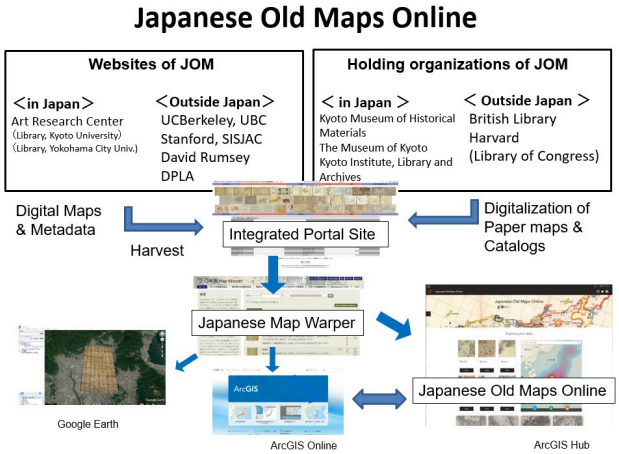


Figure 1
Framework of Japanese Old Maps Online

Measuring the Use of Tools and Software in the Digital Humanities: A Machine-Learning Approach for Extracting Software Mentions from Scholarly Articles

Zarei, Alireza
alireza.zarei@gwdg.de
GWDG

Seung-Bin, Yim
Seung-Bin.Yim@oeaw.ac.at
Austrian Centre for Digital Humanities and Cultural Heritage

Fischer, Frank
frank.fischer@dariah.eu
DARIAH-EU

Ďurčo, Matej
matej.durco@oeaw.ac.at
Austrian Centre for Digital Humanities and Cultural Heritage

Wieder, Philipp
philipp.wieder@gwdg.de
GWDG

Introduction

Tools and software are an important part of Digital Humanities (DH) practice (Dombrowski 2014, Barbot et al. 2019, Barbot et al. 2020, Fischer/Moranville 2020, Dombrowski 2021, Fischer et al. 2021, Luhmann/Burghardt 2021). Previous attempts to gain an overview of these tools were mainly based on manual aggregations, as in the case of the long-running Canadian project TAPoR.[1]Around 1,500 tools can be found there, an order of magnitude that should illustrate how difficult it is to keep everything up to date. To learn more about the actual use of tools in scientific work, especially in the Digital Humanities, we present a machine-learning approach for extracting tools and software mentioned by name in scientific publications, adding to other recent endeavours in this field (Du et al. 2020, Henny-Krahmer/Jettka 2022).

Related Works

Different approaches have emerged over the years for named entity recognition (NER), which can be categorised broadly into two groups, rule-based and machine-learning (ML) approaches (Isozaki 2002).

Rule-based approach

One of the main methods used for extracting facts, including named entities, from texts, are Hearst patterns (Hearst 1992). Hearst patterns were revisited by the Facebook team for their Hypernymy Suite tool (Roller et al. 2008) and were still found to be superior in accurately extracting relations compared to other distributional methods. Similar efforts include GATE (Cunningham et al. 2013)and other legacy solutions (Chiticariu et al. 2013).

Machine-learning approach

Among the relevant methods that use ML, we looked at SoftwareKG (Schindler et al. 2020),which uses DBpedia (Lehmann et al. 2015) to validate entities as software. The results, though, lack various software mentions from other sources, because of memorising the words (memorisation effect) (Arpit et al. 2017). Another example is the GROBID

tool[2](Du et al. 2020). After applying it to selected DH publications, we found that the recall is limited and that there is some memorisation effect. Given the recent efforts in this area (see also Henny-Krahmer/Jettka 2022), we were motivated to see if we could solve these issues to get a model that is production-ready.

Our Approach

Following the ML-based approach and considering the limitations of related work, we started to build a model that can recognise a tool based on its context (e.g. neighbourhood and grammar) and appearance (e.g. capitalised words, adjacent numbers, etc.).

Dataset

For the dataset, we preprocessed a deduplicated collection of sentences from PLOS Sociology, Linguistics and abstracts from ADHO's annual DH conferences (2015 and 2020), resulting in 1,899,652 sentences.

We created two versions of the dataset. The baseline dataset was created by preparing approximately 55,000 tool mentions[3]as patterns by processing names of tools and software coming from TAPoR and Wikidata. These patterns were fed into Prodigy,which suggested 2,205 sentences containing tool names. Following the annotation guidelines,[4]1,000 of them were annotated manually using Prodigy as the baseline dataset.[5]

In order to avoid the memorisation of tool names in our patterns, we conducted another round of annotations using Prodigy's manual-annotation feature with suggestions from a model. The suggestions of the baseline model were corrected using Prodigy, focusing on false-positive and false-negative suggestions, resulting in an additional 583 high quality-annotations as our second dataset as corrections.

Model training

Four different models were trained and evaluated to find the best training strategy for the context of the task. All models were trained using compounding batch size, a drop-out rate of 0.2 and a split of the evaluation set of 0.2 over 20 iterations.

Transfer learning

As shown in related work (Ruder et al. 2019), we wanted to see if transfer learning would improve our results. Two models, based on the first two models, were trained with transfer learning by pre-training spaCy's[6] en_vectors_web_lg model on our entire corpus of sentences.

Results

The performance of the three trained models is shown in Table 1.

Table 1: Results of different Tool Entity recognition models.

Model	Precision	Recall	F-Score
Baseline	.89	.83	.86
Baseline with corrections	.90	.88	.89
Baseline with Transfer Learning	.89	.84	.86
Baseline with corrections & Transfer Learning	.91	.92	.92

After training the corrected model, it can be seen that the model has improved significantly, especially regarding recall. This was a major criterion where many other models failed as a result of the memorisation effect (Arpit et al. 2017) of their selected tools or software. Adding newly found tools that were not present in our 55,000 tool examples and fixing the errors of our first model contributed to this result. The limitations of using a single task NLP model with a single dataset have already been studied (Ruder et al. 2019), and most recent practices consider transfer learning for their solutions. While applying transfer learning on the baseline without corrections showed no significant improvements, it significantly improved the F-score when applied together with corrections.

Application of the model to real data

The trained NER model was used to extract tool names from publications already ingested in the SSHOC Marketplace (Zarei et al. 2022). 470 publications, consisting of 54,841 sentences, were fed into the model. 2,257 different potential tool names were suggested by the NER model from 5,091 sentences mentioning tools.

Evaluation of extracted tool names

Since the discovery of previously unseen tool names is the most interesting benefit of the NER model, the evaluation of suggested tool names is important. Suggested tool names were evaluated semi-automatically by looking them up in the Marketplace and Wikidata.From the 2,257 distinct tool name suggestions, 125 were available in Marketplace entries and 38 were available in Wikidata. The rest of the suggestions were evaluated manually.

Future Work

Exploring the use of Transformer models

Transformer architecture based models such as BERT (Devlin et al. 2019) have given better results for many downstream tasks, including named entity recognition. It will be interesting to fit such transformer models to the task described in this paper and compare the results.

Validation of the model on real data

In order to monitor the performance of the NER model and detect its decay, it is important to design a feasible evaluation step that includes an automatic lookup in external resources and a manual curation to trigger the retraining of the model.

Bibliography

Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S. et al .(2017). A Closer Look at Memorization in Deep Networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML 2017), pp. 233–242.

Barbot, L., Fischer, F., Moranville, Y., Pozdniakov, I.(2019): Which DH Tools Are Actually Used in Research? In: weltliteratur.net, 6 December 2019. (URL: https://weltliteratur.net/dh-tools-used-in-research/)

Barbot, L., Dombrowski, Q., Fischer, F., Rockwell, G., Spiro, L.(2020): Who Needs Tool Directories? A Forum on Sustaining Discovery Portals Large and Small. In: DH2020: "carrefours/intersections". 22–24 July 2020. Book of Abstracts. University of Ottawa.

Chiticariu, L., Li, Y., & Reiss, F. R.(2013). Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems! In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 827–832.

Cunningham, H., Tablan, V., Roberts, A., Bontcheva, K.(2013). Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. PLOS Computational Biology 9(2), doi:10.1371/journal.pcbi.1002854.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.(2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. doi:10.48550/arXiv.1810.04805.

Dombrowski, Q.(2014): What Ever Happened to Project Bamboo? In: Literary and Linguistic Computing, Vol. 29, Issue 3, September 2014, pp. 326–339, doi:10.1093/llc/fqu026.

Dombrowski, Q.(2021): "The Directory Paradox." In: Anne McGrail et al. (eds.): Debates in Digital Humanities: Institutions, Infrastructures at the Interstices. University of Minnesota Press, pp. 83–98.

Du, C., Howison, J., Lopez, P. (2020). Softcite: Automatic Extraction of Software Mentions in Research Literature. Poster contribution. 1st SciNLP workshop at AKBC.

Fischer, F., Moranville, Y.(2020): "DH Tools Mentioned in 'The Programming Historian'." In: weltliteratur.net, 17 Jan 2020. (URL: https://weltliteratur.net/dh-tools-programming-historian/)

Fischer, F., Burghardt, M., Luhmann, J., Barbot, L., Moranville, Y., Zarei, A.(2021): Die Werkbänke der Digital Humanities: Zur Rolle von Tools und Software für die Forschungsarbeit. In: vDHd2021: "Experimente", Zenodo, doi:10.5281/zenodo.4639228.

Hearst, M.A.(1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In: COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics.

Henny-Krahmer, U., Jettka, D.(2022). Softwarezitation als Technik der Wissenschaftskultur: Vom Umgang mit Forschungssoftware in den Digital Humanities. DHd2022: Kulturen des digitalen Gedächtnisses. 7–11 March 2022. Book of Abstracts. University of Potsdam, doi:10.5281/zenodo.6328047.

Isozaki, H., Kazawa, H.(2002). Efficient support vector classifiers for named entity recognition. In: Proceedings of the 19th International Conference on Computational Linguistics. Volume 1, pp. 1–7, doi:10.3115/1072228.1072282.

Lehmann, J. et al. (2015). DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. In: Semantic Web 6(2), pp. 167–195.

Luhmann, J., Burghardt, M.(2021): Digital humanities – A discipline in its own right? An analysis of the role and position of digital humanities in the academic landscape. In: Journal of the Association for Information Science and Technology, pp. 1–24, doi:10.1002/asi.24533.

Roller, S., Kiela, D., & Nickel, M.(2018). Hearst patterns revisited: Automatic hypernym detection from large text corpora, doi:10.48550/arXiv.1806.03191.

Ruder, S., Peters, M.E., Swayamdipta, S., & Wolf, T. (2019, June). Transfer learning in natural language processing. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials (pp. 15–18).

Schindler D., Zapilko B., Krüger F.(2020). Investigating Software Usage in the Social Sciences: A Knowledge Graph Approach. In: Harth A. et al. (eds.): The Semantic Web. ESWC 2020. Lecture Notes in Computer Science, vol 12123. Springer, Cham, doi:10.1007/978-3-030-49461-2_16.

Zarei, A., Seung-Bin, Y., Ďurčo, M., Illmayer, K., Barbot, L., Fischer, F., Gray, E. (2022). Der SSH Open Marketplace: Kontextualisiertes Praxiswissen für die Digital Humanities. In: DHd2022: "Kulturen des digitalen Gedächtnisses". 7–11 March 2022. Book of Abstracts. University of Potsdam, doi:10.5281/zenodo.6327975.

Notes

1. https://tapor.ca/
2. https://cloud.science-miner.com/software/
3. https://gitlab.gwdg.de/sshoc/data-ingestion/-/blob/master/repositories/extraction/data/patterns/all_final.jsonl
4. https://gitlab.gwdg.de/sshoc/data-ingestion/-/blob/master/Annotation%20Guideline%20for%20Tool%20Extraction.pdf
5. https://gitlab.gwdg.de/sshoc/data-ingestion/-/tree/master/repositories/extraction/annotation/result
6. https://spacy.io/, v2.3.2

Transfer Learning for Olfactory Object Detection

Zinnen, Mathias
mathias.zinnen@fau.de
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

Madhu, Prathmesh
prathmesh.madhu@fau.de
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

Bell, Peter
peter.bell@uni-marburg.de
Philipps-Universität Marburg, Germany

Maier, Andreas
andreas.maier@fau.de
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

Christlein, Vincent
vincent.christlein@fau.de
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

Introduction

Smells are an important, yet overlooked part of cultural heritage (Bembibre and Strlič, 2017). TheOdeuropa project [1] analyzes large amounts of visual and textual corpora toinvestigate the cultural dimensions of smell in 16th – 20th century Europe.The study of pictorial representations bears a specific challenge: the substrate of smell is usually invisible (Marx, 2021). Object detection is a well-researched computer vision technique, and sowe start with the recognition of objects, which may then serve as a basisfor the indirect recognition of more complex, and possibly more meaningful,smell references such as gestures, spaces, or iconographic allusions (Zinnen, 2021).



Figure 1:
Smell. The Five Senses. 1558 – 1617. Jan Pietersz Saenredam. National Gallery of Art.

However, the detection of olfactory objects in historical artworks is achallenging task. The visual representation of objects differs significantlybetween artworks and photographs (Hall et al., 2015). Since state-of-the-art object detectionalgorithms are trained and evaluated on large-scale photographic datasetssuch as ImageNet (Russakovsky et al., 2015), MS COCO (Lin et al., 2014), or OpenImages (Kuznetsova et al., 2020), their performancedrops significantly when applied to artistic data. This domain gap betweenstandard object detection datasets and artistic imagery can be mitigatedby training directly on artworks, either by using existing datasets or bycreating an annotated dataset for the target domain. Another challengeis the mismatch between object categories present in modern datasets andhistorical olfactory objects, caused by historical diachrony on the one hand (Marinescu et al., 2020), and the particularity of some smell-relevant objects on the other (Ehrich et al., 2021).

Methodology

To overcome the domain gap and category mismatch between our applicationand the existing datasets, we apply transfer learning – a training strategy where machine learning algorithms are pre-trained in one domain and then fine-tuned in another, greatly decreasing the amount of required training data in the target domain (Sinno and Yang, 2009; Zhuan et al., 2020, Madhu et al., 2020).
We are continuously collecting and annotating artworks with possible olfactory relevance from multiple museum collections. Based on these, we created a dataset of olfactory artworks containing 16728 annotations on 2229 artworks. From this full set of annotations, we created a test set of 3416 annotations on 473 artworks, while the remaining data was used for training.

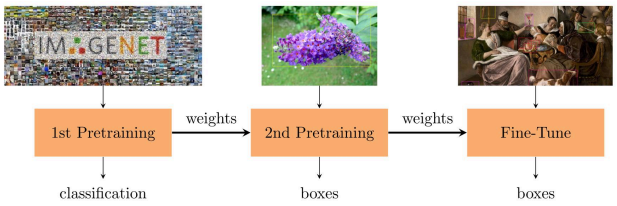


Figure 2:
Category overlap between Odeuropa & OpenImages categories

A common transfer learning procedure is to use detection backbones that have been pre-trained on ImageNet and fine-tune them for object detection (Zhuang et al., 2020). We expand this strategy by an additional pre-training step, where we train an ImageNet pre-trained object detection network (Ren et al., 2015) using different datasets. Finally, we fine-tune the resulting model using our olfactory artworks dataset (fig. 3).

For pre-training, we use three

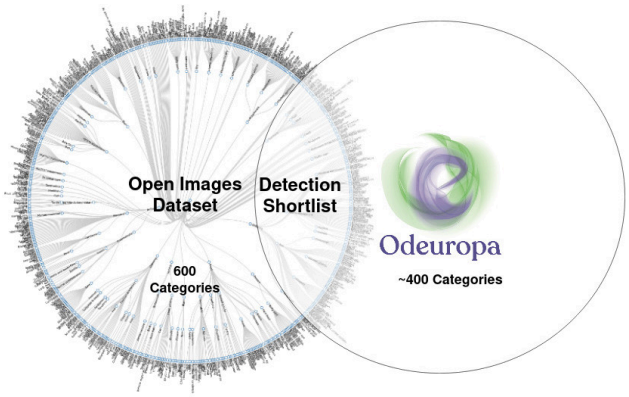


Figure 3:

Category overlap between Odeuropa & OpenImages categories

different datasets, deviating to varying amounts from our olfactory artworks dataset in terms of categories and style (table 1): a) Same Categories, Different Styles - A subset of OpenImages (OI) containing only odor objects results in a complete category match (fig. 2); however, since OpenImages contains only photographs, there is a considerable style difference. b) Different Categories, Same Styles - We apply two object detection datasets from the art domain, which are more similar in terms of style but contain different object categories, namely IconArt (IA,Gonthier et al., 2018) and PeopleArt (PA,Westlake et al., 2016).

Table 2: Evaluation of object detection performance. The best performing model pre-trained with OI achieves an improvement of 6.5% pascal VOC mAP, and 3.4% COCO mAP over the baseline method without intermediate training. We report the evaluation for each pre-training dataset, averaged over five models, fine-tuned for 50 epochs on our olfactory artworks datasets. Best evaluation results are highlighted in bold. The merge of two datasets D1 and D2 is written as D1 ∪ D2.

Dataset	Domain Similarity	Category Similarity	# Categories
OpenImages (OI)	Low	Complete match	29
IconArt (IA)	High	Medium	10
PeopleArt (PA)	Medium	Low	1

Results

To ensure a fair comparison between the different pre-training datasets, we reduce each of the datasets to the same size, train three models, and select the best according to a fixed validation set for each dataset. Additionally, we merge all three datasets, i. e., combining OI, IA, and PA, using the union over their respective classes. The resulting models are then fine-tuned on the

training set of the olfactory artworks dataset and evaluated on a separate test set. To mitigate random variations that can occur during the training process, we train five separate models for each experimental setting and report their average. Evaluation results are reported in pascal VOC (mAP 50, Everingham et al., 2010) and COCO mAP (mAP 50:95:5 (Lin et al., 2014), the two standard metrics to evaluate object detection models. We conduct two separate sets of experiments: In the first, we fine-tune the whole network, including the backbone, to assess the detection performance under realistic conditions (table 2).

Table 3: Evaluation of object detection performance for fine-tuning of thedetection heads only. All pre-training schemes increase the detection perfor-mance, while pre-training with OI leads to the best results with an increaseof 7.7% mAP 50 or 4% COCO mAP. For every pre-training dataset, we report the evaluation averaged over five models, fine-tuned for 50 epochs onour olfactory artworks datasets each. Best evaluation results are marked in bold. The merge of two datasets D1 and D2 is written as D1 ∪ D2.

Pretraining Dataset	Pascal mAP (%)	COCO mAP (%)
None (Baseline)	16.8 (±1.3)	8.4 (±0.4)
OI	**23.3** (±0.5)	**11 .8** (±0.4)
IA	22.6 (±1.2)	10.9 (±0.9)
PA	21.9 (±0.4)	10.5 (±0.2)
IA∪OI	21.8 (±0.1)	10.5 (±0.3)
IA∪PA	22.0 (±0.8)	10.6 (±0.3)
PA∪OI	22.6 (±0.3)	10.8 (±0.2)
OI∪IA∪PA	21.8 (±0.4)	10.5 (±0.2)



Figure 4:
Exemplary object predictions for a detection model without intermediate training (left), with PeopleArt pretraining (middle), and ground truth bounding boxes (right). Painting: Boy holding a pewter tankard, by a still life of a duck, cheeses, bread and a herring. 1625 – 1674. Gerard van Honthorst. RKD Digital Collection (https://rkd.nl/explore/images/287165).

We observe a performance increase for all used pre-training datasets, with an increase of 6.5%/3,4% boost in mAP 50 and COCO mAP, respectively, for the best performing pre-training scheme, which was achieved using the OI dataset. The exemplary object predictions in fig. 4 show that adding an additional pre-training stage can increase the number of recognized objects. In a second set of experiments, we train only the detection head while the backbone remains frozen, to compare the quality of the intermediate representations that have been learned using the different pre-training schemes (table 3).

Pretraining Dataset	Pascal mAP (%)	COCO mAP (%)
None (Baseline)	11.7 (±0.2)	5.5 (±0.1)
OI	**19.4** (±0.3)	**9.5** (±0.1)
IA	13.8 (±0.4)	6.4 (±0.2)

PA	13.5 (±0.2)	6.7 (±0.1)
IA∪OI	16.0 (±0.3)	7.4 (±0.2)
IA∪PA	14.6 (±1.0)	6.7 (±0.5)
PA∪OI	15.8 (±0.7)	7.3 (±0.4)
OI∪IA∪PA	16.4 (±0.6)	7.6 (±0.2)

While all pre-training schemes increase the performance, the relative increase for the OI dataset is remarkably higher. This suggests that the style similarity between the IA and PA datasets and our target dataset is less important than we expected. We can not yet conclude whether the superior performance of the OI dataset is due to the similarity in target categories. It could also be caused by other properties of the dataset. Further ablations, e. g., varying the set of OI categories are needed to more precisely assess the impact of category similarity on the detection performance, which we plan to conduct in a follow-up study. Interestingly, the performance of the merged datasets increases even in cases where OI is not part of the dataset merge. Given that we did not apply a sophisticated merging strategy,the performance increase for training with merged datasets is encouraging. Developing strategies to improve the consistency of the merged dataset, e. g., weak labeling of categories not present in the respective merge partners, represents another promising line of future research. We conclude that including an additional stage of object-detection pre-training can lead to a considerable increase in detection performance. While our experiments suggest that style similarities between pre-training and target dataset are less important than matching categories, further experiments are needed to verify this hypothesis.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004469

Bibliography

Bembibre, C. and Strlič, M. (2017). Smell of heritage: a framework for the identification, analysis and archival of historic odours. *Heritage Science*, **5**(2): 1–11.

Ehrich, S. C., Verbeek, C., Zinnen, M., Marx, L., Bembibre, C. and Leemans, I. (2021). Nose-First. Towards an Olfactory Gaze for Digital Art History. *2021 Workshops and Tutorials-Language Data and Knowledge, LDK 2021*. pp. 1–17.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J. and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, **88**(2). Springer: 303–38.

Gonthier, N., Gousseau, Y., Ladjal, S. and Bonfait, O. (2019). Weakly Supervised Object Detection in Artworks. In Leal-Taixé, L. and Roth, S. (eds), *Computer Vision – ECCV 2018 Workshops*, vol. 11130. (Lecture Notes in Computer Science). Cham: Springer International Publishing, pp. 692–709 doi: 10.1007/978-3-030-11012-3_53. http://link.springer.com/10.1007/978-3-030-11012-3_53 (accessed 11 April 2022).

Hall, P., Cai, H., Wu, Q. and Corradi, T. (2015). Cross-depiction problem: Recognition and synthesis of photographs and artwork. *Computational Visual Media*, **1**(2). Springer: 91–103.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., et al. (2020). The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, **128**(7): 1956–81 doi: 10.1007/s11263-020-01316-z.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In Fleet, D., Pajdla, T., Schiele, B. and Tuytelaars, T. (eds), *Computer Vision – ECCV 2014*, vol. 8693. (Lecture Notes in Computer Science). Cham: Springer International Publishing, pp. 740–55 doi: 10.1007/978-3-319-10602-1_48. http://link.springer.com/10.1007/978-3-319-10602-1_48 (accessed 13 September 2021).

Madhu, P., Villar-Corrales, A., Kosti, R., Bendschus, T., Reinhardt, C., Bell, P., Maier, A. and Christlein, V. (2020). Enhancing human pose estimation in ancient vase paintings via perceptually-grounded style transfer learning. *ArXiv Preprint ArXiv:2012.05616*.

Marinescu, M.-C., Reshetnikov, A. and Lopez, J. M. (2020). Improving object detection in paintings based on time contexts. *2020 International Conference on Data Mining Workshops (ICDMW)*. Sorrento, Italy: IEEE, pp. 926–32 doi: 10.1109/ICDMW51313.2020.00133. https://ieeexplore.ieee.org/document/9346513/ (accessed 9 June 2021).

Marx, Lizzie (2021). Perfume and books of secret. *Exhibition Catalogue Mauritshuis,* The Hague.

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, **22**(10). IEEE: 1345–59.

Ren, S., He, K., Girshick, R. and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, **28**.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., et al. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, **115**(3): 211–52 doi: 10.1007/s11263-015-0816-y.

Westlake, N., Cai, H. and Hall, P. (2016). Detecting people in artwork with cnns. *European Conference on Computer Vision*. Springer, pp. 825–41.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H. and He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, **109**(1). IEEE: 43–76.

Zinnen, M. (2021). How to See Smells: Extracting Olfactory References from Artworks. *Companion Proceedings of the Web Conference 2021*. Ljubljana Slovenia: ACM, pp. 725–26 doi: 10.1145/3442442.3453710. https://dl.acm.org/doi/10.1145/3442442.3453710(accessed 16 September 2021).

Notes

1. https://odeuropa.eu

Short Presentations

Analysis of the Gutenberg 42-line Bible types aided by type-image recognition

Agata, Mari
agatamari@keio.jp
Keio University, Japan

Agata, Teru
agata@asia-u.ac.jp
Asia University, Japan

- Background and purpose

While the traditional view of the type casting method used at the earliest stage of European printing asserts that identically shaped types were produced by metal punch, matrix, and hand mould, this has been under renewed debated over the past two decades. Reported in 2000, Paul Needham and Blaise Agüera y Arcas's clustering analysis of images of the lowercase "i" of the Donatus Kalender type (DK type), that is, Johann Gutenberg's first type, discovered hundreds of clusters of "i," leading to the conclusion that "[e]ither many matrices were used in parallel, or equivalently, the matrix was temporary and needed to be re-formed between castings, – or both" (Agüera y Arcas, 2003). Despite the considerable attention their research attracted, there have been relatively a few substantial follow-up studies. One of them is a clustering analysis of small samples – ten "i"s and 21 "a"s on a single page of the Gutenberg 42-line Bible (the B42) printed around 1455–, which corresponds to the conclusion of the DK type analysis (Alabert and Rangel, 2011).

The authors have tried to contribute to this argument by analyzing the B42 types. The typeface is Gothic Textura and very similar to the DK type, but smaller in size. Nearly 300 types have been identified by earlier scholars, due to many variations of each letter. This paper is an interim report of the ongoing type analyses of the B42.

- Method

The digital images of the Keio University Library copy of the B42 (vol. 1 only) were used for analyses. Type-image recognition reinforced by machine learning was executed with an open-source OCR engine, Tessaract-OCR, followed by manual corrections. Each piece of type-image data had information about X and Y coordinates, pixel width and height, and transcribed character. Type-image recognition have been completed, 46 pages of which containing around 120,000 letters have been manually corrected and used for analyses.

The first analysis is to calculate the vertical distance between a suspension stroke to show suspension of nasal "n," "m," and other letters and companion letter. The authors previously made a statistical analysis of a single letter "ū" that appeared on selected 42-line pages. The results showed that the variation was too wide for letters cast from a single matrix, thereby suggesting that they were made from multiple matrices (Agata and Agata, 2021). Several other types that appeared on 42-line pages as well as on 40-line pages, each with a suspension stroke, are newly analyzed. The 40-line pages had been printed at the very earliest stage of the print run, before the lines per page were increased to 42, thus enabling a chronological analysis.

The second method is to analyze the horizontal position of suspension strokes in relation to a companion letter. The widths of suspension strokes and companion letters, and the distance of their horizontal centers are calculated based on the brightness of the pixels.

The method of identifying inverted letters is also explored. The letters "n" and "u" are printed upside down in some places, thus an inverted "n" looks like "u", and *vice versa*. The similarity of "u"s are measured by contour extraction, then the resultant similarity matrix is used for clustering analysis.

- Results

The results of vertical distance analysis show that the distribution of the distance between a suspension stroke and companion letter of the types used in 40-line pages is limited to a narrower range than that of 42-line pages. This may suggest that matrices were added during the print run.

The horizontal position analysis indicates that the scatter diagram shows no correlation between the relative position of strokes and companion letters and that the strokes and letters are more likely to have been punched separately rather than together with whole-letter punches.

The clustering analysis of "u"s shows the potential usefulness of the method together with several challenges.

It should be noted that a relatively small number of types have been analyzed, but the analyses so far are promising. The present methods are not limited to the analysis of the B42, but are applicable to any early printed books. A culmination of examples will contribute to further discussions on type casting methods in European printing.

- Acknowledgements

The B42 images taken by the HUMI Project were provided for research and reproduced by courtesy of the

Keio University Library. This study was supported by JSPS KAKENHI Grant Number 18H03496.

Bibliography

Agata, M. Agata, T. (2021). Statistical analysis of the Gutenberg 42-line Bible types: special focus on letters with a suspension stroke. *The Papers of the Bibliographical Society of America*, 115 (2): 167-83, 10.1086/713981.

Agüera y Arcas, B. (2003). Temporary matrices and elemental punches in Gutenberg's DK type. In Jensen, K. (ed), *Incunabula and Their Readers: Printing, Selling and Using Books in the Fifteenth Century*. London: British Library, pp. 1–12.

Alabert, A. and Rangel, L. M. (2011). Classifying the Typefaces of the Gutenberg 42-line Bible. International Journal on Document Analysis and Recognition, 14: 303–17, 10.1007/s10032-010-0140-6.

RDF-star-based Digital Edition of Travel Journals

Alassi, Sepideh
sepideh.alassi@unibas.ch
DHLab, University of Basel, Switzerland

Rosenthaler, Lukas
lukas.rosenthaler@unibas.ch
DHLab, University of Basel, Switzerland

Our project aims to develop tools and infrastructure for the creation of interactive web-based digital editions of metadata oriented-documents such as travel journals based on RDF-star and SPARQL-star. [1] Successful digital editions have been created as RDF-based knowledge graphs enabling users to study the editions as a network of interconnected resources. Standard RDF principles are used to define ontologies for modeling metadata and textual information of these editions as RDF triples. The platforms presenting these editions use standard SPARQL for data analysis and query. Standard RDF, however, is not an optimal choice for the digital edition of metadata-oriented documents such as travel journals because most of the information in such documents is accompanied by metadata information describing it, e.g. "person A was at location B" for a certain period of time. Creating statements about statements using standard RDF is troublesome. The very first RDF 1.0 specification uses the mechanism called reification for supporting statements about statements. Reification, however, introduces processing overhead due to the increased number of additional statements needed to identify the reference triple and appears too verbose when represented in RDF and SPARQL (Kasenchak et al. 2021). RDF-star and SPARQL-star overcome this deficit with an extension of the RDF standard and increase the efficiency of queries by reducing the query time. RDF-star allows for triples that represent metadata about another triple by directly using this other triple as its subject or object (Hartig 2017). Using RDF-star, we can easily attach metadata to the edges of the knowledge graph that represents the metadata-oriented document. Our infrastructure will provide tools based on SPARQL-star to efficiently query the data.

The technical basis for our project is Knora, [2] an infrastructure for humanities data consisting of an RDF-triplestore, an OWL base ontology, and a RestFul API that allows for querying and adding to the data. For our project, this infrastructure will be further developed to support RDF-star and SPARQL-star. As a prototype document to use for developing the ontologies, tools, and the infrastructure, we have chosen Jacob Bernoulli's travel diary. Jacob Bernoulli (1654–1705) was the first mathematician of the Bernoulli dynasty who, like many in his time, traveled in pursuit of knowledge. He kept a record of his trips in a small journal called Reisbüchlein from August 1676 until October 1683 when he permanently settled in Basel. The entries of this journal contain brief descriptions of places he visited, people he met, travel costs, and the events and phenomena he witnessed during his trips. This so far unresearched document is kept at the archive of University of Basel. Our project aims at creating an open-access RDF-star based edition of this document making every piece of information within it efficiently queryable. Jacob Bernoulli's scientific notebook Meditationes is currently available on the BEOL[3] platform as an RDF-based digital edition together with the digital edition of correspondence of members of Bernoulli dynasty and Leonhard Euler (Schweizer, Alassi 2018). The digital edition of Reisbüchlein will be integrated into this platform allowing researchers to follow Bernoulli's line of thoughts from his travel diaries to the scientific ideas written in Meditationes at the same time and his correspondences. Based on this document, we will develop a generic RDF-star-based ontology describing textual data and metadata of travel diaries.

To create a normalized edition of Reisbüchlein whose text is written in old German and French, we have chosen a semi-automatic approach. There is an old unpublished typed transcription of this journal available which we employed to generate digital annotations using Transkribus. [4] An editor is currently verifying the automatically generated annotations consulting the digitized facsimiles. At the same time, editorial commentaries regarding the structure of the text, content, and explanation of specific terms are being added to the annotations. Through Knora API,

interlinked resources will be created for image regions, their annotations, and commentaries.

We intend to use NLP algorithms to automatically recognize and tag the named entities within the text, such as locations and persons. The tagged entities will then be verified by comparing against the glossary given in the old transcription that lists the places and people mentioned in Reisbüchlein. The algorithm will then find (by querying Wikidata) and add geo-identifiers to locations and GND numbers to persons and will create resources for locations and persons. The tagged elements within the text will be linked to the corresponding resources. This will allow queries for a text that contains a certain person and/or location. [5]

Bibliography

Hartig, Olaf. "RDF* and SPARQL*: An Alternative Approach to Annotate Statements in RDF". International Semantic Web Conference 2017.

Schweizer, T. and Alassi, S. (2018) "Bernoulli-Euler Online: Development of a Platform for Early Modern Mathematical Texts as Part of a Generic Infrastructure", in *Digital Humanities Congress 2018*. Sheffield: Lana Pitcher and Michael Pidd. Proceedings of the Digital Humanities Congress 2018. Studies in the Digital Humanities, pp. 1–4. Available at: https://www.dhi.ac.uk/openbook/chapter/dhc2018-schweizer.

Kasenchak, Bob, Aren Lehnert and Gene Loh, "Use Case: Ontologies and RDF-Star for Knowledge Management". *The Semantic Web: ESWC 2021 Satellite Events, LNCS 12739*, 2021, 254–260. https://doi.org/10.1007/978-3-030-80418-3_38.

Notes

1. https://www.ontotext.com/knowledgehub/fundamentals/what-is-rdf-star/
2. https://dsp.dasch.swiss/
3. https://beol.dasch.swiss/
4. https://readcoop.eu/transkribus/
5. See the proof of concept in "open research data queriable by location" report of Swiss ORD hackathon 2021, in https://docs.google.com/document/d/1lbD6go_mSNAH3Gmj_Ao9GFnGwPFd4ZyGo1HrP_tYtBs/edit#

On Epistemic Comparability and Challenges on Data Reuse:The Experience of READ-IT

Antonini, Alessio
alessio.antonini@open.ac.uk
The Open University, United Kingdom

The EU JPI Reading Europe Advanced-Data Investigation Tools (READ-IT) project (https://readit-project.eu) proposed a novel approach to a pan-European research agenda on reading based on shared tools and data interoperability "by design" (Our-Vial & Antonini, 2021). These tools were developed in parallel with and support a wide diverse range of case studies (Vignale et al, 2019). The toolkit includes a) the Reading Experience Ontology (REO) (Vignale et al, 2020; Antonini et al, 2021), b) the text-annotation tool (https://read-it.hum.uu.nl) and c) the platform for managing reading testimonies (https://read-it.in-two.com). However, while reflecting on using the data at scale, a gap in the models emerged concerning the epistemology of sources, i.e. the conditions about testimonies of reading were generated. This contribution reports on this gap, its implications for future visions for research interoperability and our approach to address it.

The focus of READ-IT on data interoperability led to good results toward a shared definition of the phenomena and response to reading. However, the information needed to interpret testimonies extends beyond readers and reading, e.g., to the situation that led to the generation of the testimony, the socio-economical context, and the temporal distance between reading and the testimony. The result is an epistemological barrier to data reuse, comparability, and integration of experiential studies. The following two real examples expose the hidden implications of building new reading case studies on existing data from, e.g., linked open data service.

Case 1 - Exploration and reuse. The discovery and available data on reading could exploit either provenance or metadata or, more specifically, aspects of the reading experience. For instance, PoKUS (https://pokus.ffzg.unizg.hr/en/) – a study on reading in Croatia – aims to reuse READ-IT data about reading memories (REO concept of "Memory" specialisation of "State of Mind" [5]). The collected memory annotations [7] include evidence from WWI diaries, readers of Russian periodicals, Soviet Czech school diaries, social media, interviews, or online reading groups. While discussing with researchers about their specific case studies, it was clear that, e.g., social pressure and censorship on Czech students limited

the insight that could be extracted by their carefully crafted school diaries or soldier diaries written as part of a veteran support programme, reporting reading years after they occurred.

Case 2 – Heterogeneous comparative studies. Having a shared definition of reading enables comparative studies among heterogeneous sources. For instance, an ongoing comparative study on popular reading brings together a) long, matured reviews of books from online reading groups, with b) short impulsive comments on webcomics "issues". The design of heterogeneous comparative studies also requires careful considerations (Benatti et al, 2021) about, e.g., the practices of the audiences, maturity of the experience or type of information provided, connected by the form of prompting.

To summarise, data interoperability is a precondition, but it is not sufficient: a common language of the phenomenon does not exhaust the information needed for interpretation. Specifically, in the vision of step further toward research interoperability - as the synergy between studies in terms of agenda and findings contributing to the knowledge of a common phenomenon – a computer-readable epistemology of research case studies should be included as a part of their output.

In both cases, the data do not reflect these considerations, emerged only by talking directly with the involved researchers. To address this issue, we designed two un-planned ontologies: Experience & Observation (E&O) (https://github.com/modellingDH/odp_experience) and Profiles, Groups & Communities (PGC) (https://github.com/modellingDH/profile-group-community-odp).

E&O is currently in the READ-IT contribution platform. The first application of E&O is to document the different modalities we use in crowdsourcing of experience of reading, e.g. through postcards, webforms and chatbot conversations (https://readit-project.eu/contact/contribute-to-read-it/). Secondly, E&O describes the relations between activities and prompting of reading experiences, typical of the different sources (e.g., topics of the questions asked, time from reading). This use of E&O helped identify nine recurring patterns exposing new, unexpected similarities and differences between case studies that can be evaluated through objective metrics (Antonini et al, under review).

PGC complements the characterisation of readers introduced by REO, with the difference between reader and status of the reader at the time of reading, by introducing the missing social dimension. Specifically, PGC addresses the specific reader profiles in terms of, e.g., linguistic competencies or core values, their belonging to social bodies (religious groups, political parties) or community of practice (as reading groups or professions such as editors or scholars).

This contribution addresses the epistemic gap in how data about experiential research, which is a barrier in moving beyond data interoperability toward a conceptual integration between different research case studies. This gap is addressed by introducing two models used to explicit essential information about research case studies and sources to complement and give context to research data. Both gap and solution are discussed based on actual case studies on reading.

Bibliography

Our-Vial, B. and Antonini, A. (2021) Reconnecting with the Evolving Journey of Reading. Cultural Practice, Re-Connect, 2021, Last access September 2021 https://www.culturalpractice.org/article/reconnecting-with-the-evolving-journey-of-reading.

Vignale, F., Benatti, F. and Antonini, A. (2019). Reading in Europe - Challenge and Case Studies of READ-IT Project. In: Digital Humanities Conference 2019, 9-12 Jul 2019, Utrecht, Netherland.

Vignale, F., Antonini, A. and Gravier, G. (2020). The Reading Experience Ontology (REO): Reusing and Extending CIDOC CRM. In proceedings of DH2020. 2020, Ottawa, Ca.

Antonini, A., Suárez-Figueroa, M. C., Adamou, A., Benatti, F., Vignale, F., Gravier, G. and Lupi, L. (2021). Understanding the phenomenology of reading through modelling. Semantic Web, 12(2), 191-217.

Benatti,F., Norrick-Rühl, C. and Antonini, A. (2021) Reading Popular Culture Offline and Online: Outlining a Comparative Study of Reading Experiences Between Webcomics and Twenty-First Century Book Club Choices. EADH2021, September 2021.

Antonini, A., Adamou, A., Suárez-Figueroa, M. C. and Benatti, F. (under review) Experiential Observations: an Ontology Pattern-based Study on Capturing the Potential Content within Evidences of Experiences. Special issue on Cultural Heritage 2021, Semantic Web – Interoperability, Usability, Applicability an IOS Press Journal (under review), URL: http://semantic-web-journal.net/content/experiential-observations-ontology-pattern-based-study-capturing-potential-content-within.

Salience in Literary Texts: A Combined Approach to the Relevance of Passages

Arnold, Frederik
frederik.arnold@hu-berlin.de
Humboldt-Universität zu Berlin, Germany

Fiechter, Benjamin
fiechtbe@hu-berlin.de
Humboldt-Universität zu Berlin, Germany

Gius, Evelyn
evelyn.gius@tu-darmstadt.de
Technical University of Darmstadt, Germany

Jäschke, Robert
robert.jaeschke@hu-berlin.de
Humboldt-Universität zu Berlin, Germany

Martus, Steffen
steffen.martus@rz.hu-berlin.de
Humboldt-Universität zu Berlin, Germany

Vauth, Michael
michael.vauth@tu-darmstadt.de
Technical University of Darmstadt, Germany

A combined approach to the relevance of text passages

In this contribution, we want to outline insights that arise from combining two distinct approaches to literary texts that analyse the relevance of specific text passages. We have been working on the identification of the **narrativity** represented in literary texts as well as on the **quotation** of the texts in research to identify passages especially relevant from a hermeneutical perspective. This now allows us to explore whether the structures and patterns that are emerging from these two approaches can be related to each other in a meaningful way.

Narrativity as textual relevance criterion

Our identification of narrativity of literary prose texts is based on the annotation of events. By considering the core features of events in narrative theory (i.e., being a state, a process in time and change of state) we classify each verbal phrase in a text as change of state, process event, stative event or non-event. [1] , [2] To enable measuring narrativity, this categorical scaling is transposed into a numerical scaling reflecting the degree of narrativity of the event categories. In accordance with a narrative theory understanding of events, we determine the narrativity of the annotation categories with the values 7 (change of state), 5 (process event), 2 (stative event), and 0 (non-event). By additionally smoothing the narrativity value we are able to model the narrativity of a text as a graph. [3] Figure 1 shows the narrativity graph for the novella *Die Judenbuche* by Annette von Droste-Hülshoff which serves as an example for our approach.

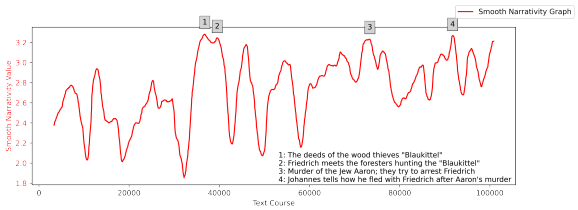


Figure 1.
Narrativity score

Key passages: quotation as textual relevance criterion

We consider key passages as parts of a literary text that are especially relevant for interpretation and can differ in length from only a few words to one or more paragraphs. To learn which parts of a text are more relevant than others, we rely on expert knowledge, which we obtain from numerous interpretations of a literary work containing quoted passages. This is a new approach in text and literature studies, that has not been theoretically founded yet; though the term “Schlüsselstelle” (key passage) and equivalents are used regularly in text interpretations in German language.[4]

For this study, we limit ourselves to a quantitative view. We have analysed 44 interpretations of *Die Judenbuche*, all in German language, published between 1995 and 2015 and identified quoted passages with a Python tool for quotation detection in fictional texts.[5] Figure 2 visualises the identified quotations over the course of the text; the histogram shows quotation frequency and the graph the smoothed frequency for each verbal phrase identified during the event annotation. Notably, the beginning and the end are quoted most frequently, together with three other frequently quoted passages.

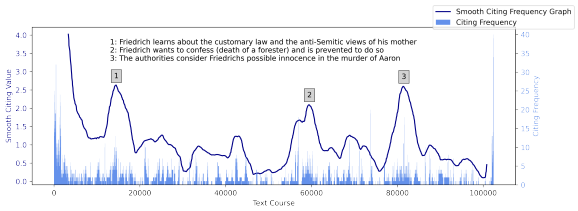


Figure 2.
Quotation frequency

Combining the approaches: frequently quoted (key) passages and narrativity

By combining exploration of narrativity and quotation frequency (cf. Figure 3) we can explore whether a passage is referred to as one that is mainly interesting for the storyline or for the interpretation. Passages with a high narrativity score are particularly important for the plot and the comprehension of the plot, while passages with a low narrativity value more often contain dialogue or narrator comments in which interpretation proposals are already made that are taken up in literary studies texts. For passages with a medium narrativity value, potential interdependencies are difficult to determine on the basis of only one text, but we are aiming to obtain more detailed knowledge on this in the future, including also non-frequency based analyses of references. Also, the beginning and the end of the text seem to be quoted in a different manner. Here quotation frequency and narrativity seem to be connected only loosely. Instead, these borders of the text seem to be used mostly to provide a framework for interpretations, in which the interpreters select the most interesting passages for their intent.

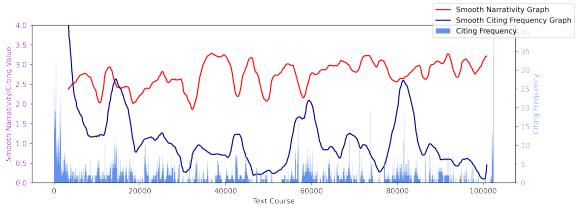


Figure 3.
Narrativity score and quotation frequency

Outlook

While these findings point out how the quotation of text passages may relate to their narrativity, they should be evaluated against a broader corpus of texts. There, the classification of functions of quotations is the most interesting aspect. We assume that plot-orientated quotations in the secondary literature correlate with higher narrativity, whereas passages quoted in order to develop a more comprehensive interpretation of the text display less narrativity. For evaluating this, we plan to combine our automated analysis of narrativity with the automated detection of key passages.

Bibliography

Arnold, F. and Fiechter, B. (2022). Lesen, was wirklich wichtig ist: Die Identifikation von Schlüsselstellen durch ein neues Instrument zur Zitatanalyse. *DHd2022: Konferenzabstracts.* https://doi.org/10.5281/zenodo.6327917 (accessed 11 April 2022).

Arnold, F. and Jäschke, R. (accepted). Lotte and Annette: A Framework for Finding and Exploring Key Passages in Literary Works. *Proceedings of the Workshop on Natural Language Processing for Digital Humanities at ICON 2021*.

Gius, E. and Vauth, M. (2022). Inter Annotator Agreement und Intersubjektivität. *DHd2022: Konferenzabstracts*. https://doi.org/10.5281/zenodo.6328209 (accessed 11 April 2022).

Vauth, M. and Gius, E. (2021). Richtlinien für die Annotation narratologischer Ereigniskonzepte. *Zenodo*. https://doi.org/10.5281/zenodo.5078174 (accessed 11 April 2022).

Vauth, M., Hatzel, H. O., Gius, E. and Biemann, C. (2021). Automated Event Annotation in Literary Texts. *CHR 2021: Computational Humanities Research Conference.* Amsterdam, pp. 333–45. http://ceur-ws.org/Vol-2989/short_paper18.pdf (accessed 11 April 2022).

Notes

1. For a detailed explanation of the manual annotation on which the data used in this contribution is based cf. Vauth and Gius, 2021.
2. For a description of classification of events as well as the automation approach and results cf. Vauth et al., 2021.
3. For a discussion of the adequacy of this implementation for literary studies, especially with regard to intersubjectivity, cf. Gius and Vauth, 2022.
4. For more details on key passages and the aim of the project cf. Arnold and Fiechter, 2022.
5. For a detailed explanation, cf. Arnold and Jäschke (accepted). Source code available at: https://scm.cms.hu-berlin.de/schluesselstellen/lotte.

Sound Predicts Meaning: Sound Iconic Relations between Vowels' Formants and Emotional Tone in German and Japanese

Auracher, Jan
auracher@nus.edu.sg
National University of Singapore; Max-Planck-Institute for Empirical Aesthetics

Menninghaus, Winfried
w.m@ae.mpg.de
Max-Planck-Institute for Empirical Aesthetics

Scharinger, Mathias
mathias.scharinger@staff.uni-marburg.de
Philipps-University Marburg

In this research project, we investigate the extent to which the phonetic properties of a text tell us something about the mood expressed in that text. Empirical research on the relationships between sound and meaning in language has provided solid evidence that the articulatory and acoustic properties of phonemes are implicitly related to non-acoustic features such as size, brightness or emotional tone (Reay, 1994; Schmidtke et al., 2014; Sidhu et al., 2018). Moreover, these sound-meaning relationships appear to be broadly universal, i.e., they are found in all languages and language families.

However, while there is a wealth of studies on phonosemantic relationships at the level of individual phonemes or words, few studies have investigated the role of sound-meaning relationships in texts (Aryani et al., 2013; Auracher et al., 2010; Fónagy, 1961; Kraxenberger & Menninghaus, 2016; Whissell, 1999). The aim of this research project is to investigate whether and to what extent the relative frequency of certain phonetic features in texts predicts the emotional mood expressed in that text. That is, we hope to develop a universal (i.e., language-independent) method for automatic sentiment analysis based on the phonetic structure of texts.

To this end, we have compiled corpora of literary texts in various languages and language families, including English, German, Spanish, Chinese, Arabic, Japanese, Thai, and other languages. For the study, these texts are phonetically transcribed (converting graphemes to phonemes) and analyzed for the relative occurrence of certain phonetic features. In experimental studies, readers will then be asked to rate the emotional tone of randomly selected texts using bipolar scales that comprise the three dimensions of evaluation, potency, and activity.

In this paper, we present preliminary results of two experiments in which we compared sound-meaning relationships in German and Japanese poetic texts. In both experiments, we examined the relationship between the formant frequencies of vowels and the emotional tone expressed in the texts. It has been reported that formant dispersion, i.e., the relative distance between first and second formants, is implicitly associated with the notion of size, strength, or dominance (Auracher, 2017; Hoshi, 2019). Our hypothesis, therefore, was that readers would be more likely to find an expression of strength, size, and dominance in texts whose average formant dispersion is relatively low than in texts with relatively high format dispersion.

In a first experiment, 42 native German speakers evaluated the emotional tone of 90 stanzas pseudo-randomly selected from a corpus of about 1400 German poems (8000 stanzas) written between 1720 and 1900. The stanzas were divided into three categories according to whether their average formant dispersion was relatively high, relatively low, or around the general average for all stanzas. Each participant rated 30 stanzas (ten per category) on six items comprising the three bipolar dimensions of Evaluation (pleasant-unpleasant), Potency (dominant-submissive) and Activity (active-passive) (Osgood et al., 1957). Linear-Mixed-Models (LMM) per EPA dimension were run to analyze the data, using stanza category as a fixed factor and participants as a random factor. In the second experiment, we repeated the experimental design with 147 undergraduates from Japan who rated 75 Tanka (a particular form of Japanese poetry) on six bipolar scales. The Tanka were written between 1868 and 1975. Each participant rated between 12 and 21 Tanka.

Our results clearly point to a significant relationship between the phonetic features of the texts on the one hand and their evaluation by the readers on the other. In particular, the ratings of items related to the dimension of Dominance showed a strong and language-independent phonosemantic effect in the predicted direction. In both languages, stanzas with a relatively low formant dispersion were rated as highly significantly more *dominant*, more *aggressive* or more *powerful* compared to stanzas with a relatively high or neutral formant dispersion. Similarly, stanzas with lower formant dispersion were rated as more *lively*, *active* or *aroused*, suggesting that low formant dispersion is associated with Activity or Arousal. In contrast, the ratings of emotional valence (pleasant-unpleasant) expressed in the texts showed no significant relationship with formant dispersion.

In this paper, we will discuss these findings in terms of the potential of sound iconicity in written language for

sentiment analysis and outline future plans for extending the research to other languages and genres. We will also present preliminary results from currently ongoing studies in which we re-analyze our data with an extended feature set. Based on these results, we will discuss the potential of using phonetic features as language-independent predictors for sentiment analysis.

Bibliography

Aryani, A., Conrad, M. and Jacobs, A.M. (2013). Extracting salient sublexical units from written texts: "Emophon," a corpus-based approach to phonological iconicity. Frontiers in Psychology, 4: e654. https://doi.org/10.3389/fpsyg.2013.00654

Auracher, J. (2017). Sound iconicity of abstract concepts: Place of articulation is implicitly associated with abstract concepts of size and social dominance. PLoS One, 12: e0187196. https://doi.org/10.1371/journal.pone.0187196

Auracher, J., Albers, S., Zhai, Y., Gareeva, G. and Stavniychuk, T. (2010). P is for happiness, N is for sadness: Universals in sound iconicity to detect emotions in poetry. Discourse Processes, 48: 1-25. https://doi.org/10.1080/01638531003674894

Fónagy, I. (1961). Communication in Poetry. WORD, 17: 194-218. https://doi.org/10.1080/00437956.1961.11659754

Hoshi H., Kwon, N., Akita, K. and Auracher, J. (2019). Semantic associations dominate over perceptual associations in vowel–size iconicity. i-Perception, 10: e2041669519861981. https://doi.org/10.1177%2F2041669519861981

Kraxenberger, M. and Menninghaus, W. (2016). Mimological reveries? Disconfirming the hypothesis of phono-emotional iconicity in poetry. Frontiers in Psychology, 7: e1779. https://doi.org/10.3389/fpsyg.2016.01779

Osgood, C. E., Suci, G. J. and Tannenbaum, P. H. (1957). The measurement of meaning. Urbana: University of Illinois Press.

Reay, I.E. (1994). Sound symbolism. In Asher, R. E. (ed.), Encyclopedia of language & linguistics. Oxford: Pergamon Press, pp. 4064–4070.

Schmidtke, D.S., Conrad, M., and Jacobs, A.M. (2014). Phonological iconicity. Frontiers in Psychology, 5: e80. https://doi.org/10.3389/fpsyg.2014.00080

Sidhu, D.M. and Pexman, P.M. (2018). Five mechanisms of sound symbolic association. Psychonomic Bulletin & Review, 25: 1619–1643. https://doi.org/10.3758/s13423-017-1361-1

Whissell, C. (1999). Phonosymbolism and the emotional nature of sounds: Evidence of the preferential use of particular phonemes in texts of differing emotional tone. Perceptual and Motor Skills, 89: 19-48. https://doi.org/10.2466%2Fpms.1999.89.1.19

The Seven Steps: Building the DiGA Thesaurus

Autiero, Serena
serena.autiero@rub.de
Ruhr University Bochum, Germany

Elwert, Frederik
frederik.elwert@rub.de
Ruhr University Bochum, Germany

Moscatelli, Cristiano
cristiano.moscatelli@rub.de
Ruhr University Bochum, Germany

Pons, Jessie
jessie.pons@rub.de
Ruhr University Bochum, Germany

DiGA, short for "Digitization of Gandharan Artefacts. A project for the preservation and the study of the Buddhist art of Pakistan", is a project that aims to digitize and catalogue a corpus of just under 2,000 Buddhist sculptures from ancient Gandhāra (Elwert and Pons 2020). Gandhāra is a historical region which covers present-day North-western Pakistan and Eastern Afghanistan and which was a pivot between South and Central Asia. Produced during the first centuries of the Common Era, the objects which the project will digitize bear testimony to the rich cultural and religious heritage of this region that was once considered a Buddhist Holy Land. [1]

Relying on a solid digital concept that complies with current standards in the field of cultural heritage, DiGA will produce a database of objects available on OpenAccess that will lend itself to exciting new research questions and will establish best practices in the field.

The DiGA Project started in February 2021 and among its very first tasks figures the development of a Thesaurus for the description of Gandharan art and – more in general – Buddhist art. Our core vocabulary has been built starting from a limited selection of sources, a prominent one being the *Repertorio Terminologico per la schedatura delle sculture dell'arte gandharica* (indicated below as 'Repertorio'), a bilingual resource in Italian and English (Faccenna and Filigenzi 2007).

In the hagiography of the Buddha we read that, right after his miraculous birth emerging from the side of his mother Maya Devi, he immediately stood up and took seven steps. [2] At this point the intended pun in the title of this presentation should be clear also to those reader not familiar with Buddhist studies. Building up the DiGA Thesaurus is part of the early life of our project, and, therefore, the coincidence of including seven steps, as shown below, appears as an auspicious sign.

Constructing the Thesaurus starting from a printed source as the Repertorio poses many challenges, it is not a simple translation from a language (analogue) to another (digital), but it is about interpreting what a language wants to convey into a completely different system.

As for the core of this presentation, the Seven Steps leading to the Thesaurus include:

1. Collecting and building upon existing sources both digital and printed.
2. Building the core of the Thesaurus starting from OCRing the Repertorio.
3. Restructure, edit and enrich the Repertorio in RDF/SKOS.
4. Add a concept hierarchy for narratives that can univocally identify scenes, episodes, actions.
5. Add a concept hierarchy for Figures including the actors involved in Gandharan narratives.
6. Reconcile the Concepts with existing entries in Getty AAT and Iconclass.
7. Work with other projects and institutions to make the DiGA Thesaurus useful beyond the project itself. [3]

These Seven Steps summarize the work done to set up what we consider an initial core to a growing tool. With the progress of digitization of unpublished material is foreseeable that it will be necessary to add more concepts; the DiGA Thesaurus is indeed a compiled resource, bound to grow with the growth of digitization of Buddhist art.

Once the Seven Steps are completed, the DiGA Thesaurus is ready to fulfil its purposes of cataloguing and research. Its use is threefold: first, it serves as a controlled vocabulary for cataloguing in the DiGA project. The hierarchical organization of the concepts improves documentation and allows to set into relation objects apparently unrelated (mostly for geographical dispersion of the collections). Second, by its open access publication as a SKOS resource, the DiGA Thesaurus serves as a best practice example in the field of Buddhist art through the introduction of controlled vocabularies for figures, monuments, objects, elements, components, etc. Finally, the DiGA Thesaurus aims to contribute to the growing use of Linked Open Data (LOD) in order to bridge different collections of Gandharan Buddhist resources, also in different media (i.e., text and visual art), and to link generic art historical resources (Getty AAT, Iconoclass, etc.) with those specific to Buddhist studies.

We are confident that with the first release of the DiGA Thesaurus we accomplished the first (seven) steps to achieve an important goal and vision for a future of Gandharan studies where the availability of digital collections will actually lead to new insights. Already during the implementation of the Thesaurus we envisioned interesting avenues for future research both on the side of Gandharan studies and on that of Digital Humanities.

Bibliography

Drachenfels, D. and Luczanits, C. (eds) (2008). 'Gandhara, The Buddhist Heritage of Pakistan, Legends, Monasteries, and Paradise'. Mayence: Philipp von Zabern.

Elwert, F. and Pons, J. (2020). 'Linked Data Methodologies in Gandhāran Buddhist Art and Texts. Pelagios Working Group Final Report', doi10.13154/rub.148.125

Faccenna, D., Filigenzi, A., and Istituto italiano per l'Africa e l'Oriente. (2007). 'Repertorio terminologico per la schedatura delle sculture dell'arte gandharica: Sulla base dei materiali provenienti dagli scavi della Missione archeologica italiana dell'IsIAO nello Swat, Pakistan'. Roma: IsIAO.

Strong, J. (2001). 'The Buddha: A short biography'. Oxford: Oneworld.

Notes

1. For an introduction to the study of Gandharan art refer to Drachenfels and Luczanits 2008.
2. For an overview on the hagiography of the historical Buddha with references to the primary sources see Strong 2001.
3. Much of the reflection around the thesaurus stems from the Pelagios Working group involving partners from the Gandhara Project, Gandhari.org, and the Italian Archaeological Mission to Pakistan.

Global and local citation networks as a new paradigm for multiple viewpoint investigation in historical literature: a case study of the Rabbinic literature corpus

Ben-Gigi, Nati
nati.bengigi@gmail.com
Bar Ilan University, Israel

Katzoff, Binyamin
binyamin.katzoff@biu.ac.il
Bar Ilan University, Israel

Schler, Jonathan
schler@gmail.com
Holon Institute of Technology, Israel

Zhitomirsky-Geffet, Maayan
maayan.zhitomirsky-geffet@biu.ac.il
Bar Ilan University, Israel

1. Introduction

Rabbinical literature is characterized by a multiplicity of viewpoints and diverse and contesting opinions rooted in earlier sources whose influence persists in subsequent essays. One of the challenges in studying literature of this kind is to identify and organize the many controversies and views and to examine their influences and development over the generations. This paper offers a new paradigm and conceptual framework for the study of viewpoint plurality in historical literature through the prism of citation networks that are embedded in it, as well as a computational methodology based on advanced machine learning algorithms for automatic citation extraction from historical texts.

The proposed methodology was applied to Responsa literature, a subfield of Rabbinical literature, as a case study. Responsa literature is a vast collection of questions and answers that discuss concrete events, in which the questioners seek the appropriate guidance from the recognized Rabbinic authorities at the given period of time. The Responsa corpus began in Iraq (central Asia) and spans over 1,300 years. In particular, we investigated to which extent the Responsa Rabbis use (and thus are influenced by) early sources from the late classical period (Talmud) and early Middle Ages (Geonim) from West and Central Asia (Iraq and the land of the Israel).

The main challenge that the system has to overcome is the variety of formats of references, that are incomplete and include numerous abbreviations. Sometimes a reference only includes the name of the author, or only the name of the book, or only a piece of text from the book appears in the citation. The problem is even more complex for historical Hebrew-Aramaic texts, such as the one in the Responsa corpus and other Rabbinic texts. Therefore, applying a rule-based approach or classic machine learning techniques as in Romanello (2016) with manually predefined features can only achieve limited recall and accuracy for this corpus, or be useful for a limited scope, period, writing style and genre, such as, the Mishnah, or 20,000 responses from the last 130 years (HaCohen-Kerner et al., 2011; Zhitomirsky-Geffet and Prebor, 2019; Waxman, 2021).

2. Research Methodology

To build an effective reference extractor for broader and heterogeneous Rabbinic corpus, this research adapts more advanced machine learning algorithms, based on Conditional Random Fields (CRF) (using MALLET library http://mallet.cs.umass.edu/ , McCallum, 2002). The system is composed from a set of five layers as follows:

- The initial layer identifies references within the given Responsa text including concatenated references or recursive references and breaks them down to a set of discrete references.
- The second layer performs part-of-reference tagging, e.g. classifies the reference words to: book name, author name, author title, or general reference words.
- The 3 rd layer handles the various nicknames, abbreviations and acronyms of an author or a book that may be used in a reference, and maps between the name found in the text and the standard name of the author/ book.
- The 4 th layer resolves ambiguity of the found names by using author and book metadata (e.g. birth/death dates and places) from the existing resources of Talmudic research.
- The last layer builds the author and book citation networks of the extracted references.
- Finally, traditional Rabbinical literature research is consulted to investigate whether, how and to what extent the viewpoints and influences of the authors are reflected in the obtained networks.

3. Results

This study's corpus comprised 5504 Responsa files from about 40 literary sources from the Rishonim ("the first ones") period dated between 11 and 15 centuries. As a

result, 557 full references (including author/s, a book title and a section/page) were identified with a precision rate of 87%. Figure 1 presents a fragment of the resulting network. As shown in Figure 2, the influence of the Babylon Talmud (and thus its approach and worldview) on the Rabbinic literature of the period of the Rishonim (1 st half of the 2 nd millennium) is significantly higher than that of the Yerushalmi Talmud (written in the land of Israel). These findings provide preliminary reflections of the diverse viewpoints of the different groups of authors (schools), as determined by the early Asian literature influence.

Although the study examined the proposed paradigm in the Responsa corpus, the proposed methodological framework can be applied to other multi-viewpoint corpora of literature in other languages, such as the Decretum, one of the essential Latin texts of Canon law, and the Digest database of Roman law.

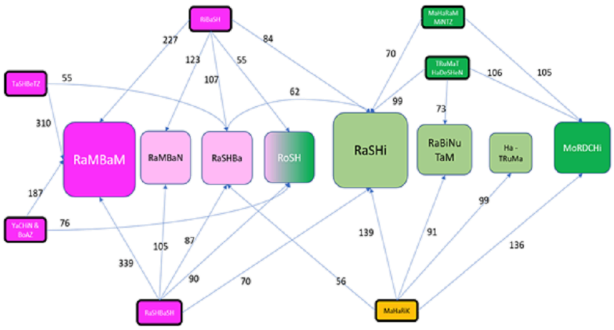


Figure 1: A fragment of the citation network based on a third of the Responsa corpus. The different colors represent various geographic regions (pink for Spain, purple for North Africa, yellow – Italy, light green – France, dark green – Ashkenaz). Rabbis with black-framed corresponding nodes are those who cite others.

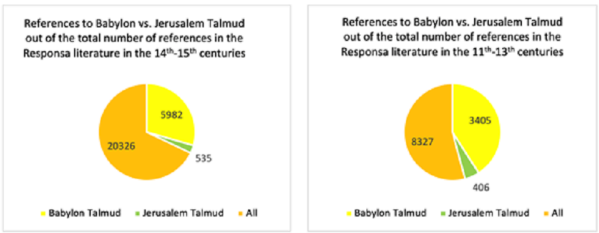


Figure 2: A relative number of references to Babylon vs. Jerusalem Talmud in the Responsa literature in different periods. As can be observed, the influence of both Talmuds' (created in Asia) on the Medieval Responsa literature decreases over time.

Bibliography

HaCohen-Kerner, Y., Schweitzer, N. and Mughaz, D. (2011). Automatically Identifying Citations in Hebrew-Aramaic Documents. *Cybernetics and Systems* 42(3): 180-197.

Romanello, M. (2016). Exploring Citation Networks to Study Intertextuality in Classics. *Digital humanities quarterly*, 10 (2).

Waxman, J. (2021). A graph database of scholastic relationships in the Babylonian Talmud, *Digital Scholarship in the Humanities*, 36(2): ii277–ii289, https://doi.org/10.1093/llc/fqab015.

Zhitomirsky-Geffet and Prebor. (2019). SageBook: Towards a cross-generational network of the Jewish sages. *Digital Scholarship in the Humanities*, 34 (3): 676-695.

A new gesture-based browsing experience for art historical research

Bernasconi, Valentine
valentine.bernasconi@uzh.ch
Digital Visual Studies, University of Zurich, Switzerland

Introduction

Thanks to the increase of digital artworks collections and the latest sophistication of computer vision techniques, there is now the possibility to computationaly apprehend large corpuses of paintings and to propose different research perspectives to the field of art history. Recent works on pattern recognition, allowing similar patterns extraction from sets of images, enable to better understand the importance of replicas within the work of an artist (Shen et al., 2019), as well as the circulation of artistic practices in the past (Lenardo, di et al., 2016). Such techniques fostered the creation of innovative browsing tools based on similarity detection from areas of images and dedicated to the practice of art history (Seguin, 2018). The specific task of automated body pose estimation also allows new methodologies for the comparison of gestures in corpuses of paintings and drawings (Impett, 2020). Applied more specifically to the hand, automated gesture retrieval can ease the task of the art historian to analyze an important number of hand gestures and their possible underlying meaning in relation to their context (Bell et al., 2013). Thus, hands and gestures can be characterised geometrically over large sets of paintings and considered through similarities or specific movements, enlighting their complex role in the pictorial practice(Impett and Süsstrunk, 2016). However, in spite of such developments, most navigation systems publicly available require some previous knowledge of the data and are still based on textual research [1]. Furthermore, other browsing approaches that rely on formal qualities

[2] or propose new interactive systems based on gestures recognition [3] (Derry et al., 2021) rather aim at fostering the curiosity of a non-expert audience in a context of cultural heritage preservation, than promoting significant discoveries for researchers. It is based on this need for an innovative browsing tool and to better apprehend a corpus of painted hands for a research project on computational and historical analysis of hands and gestures in Early Modern time that the Gestures for Artwork Browsing (GAB) project was developed.

A gesture-based tool

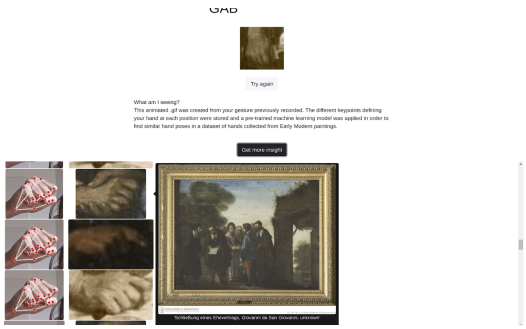


Figure 1:
Results page of the GAB application with details for each hand

The objectives of the main research are to better understand hand poses, their signification and implications in possible gestures from a dataset of Italian paintings from the 14th to the 18th Century. Based on a subset of the digitized Fototeca from the Bibliotheca Hertziana [4], a collection of 5'993 hands was created by using the *OpenPose* [5] model (Cao et al., 2021). The results of the model are a set of coordinates from keypoints describing the pose of each body detected on an image. From this information, the hands were automatically cropped.
The browsing tool consists of recording a hand gesture performed by the user in front of a webcamera for five seconds. The keypoints of the hand detected [6] at each frame are then analysed in order to retrieve, via a K-NN algorithm, similar hand poses from the collection. Finally, a .gif animation is created, reproducing the hand gesture of the user with painted hands from the Early Modern time. Apart from the animation, the user also has the possibility to see the detail of each hand used and the painting it belongs to with corresponding metadata for a better contextualisation and further investigation.

The research scope

The idea to translate a gesture to an animation has several research scopes, one of which is the possibility to show different representations of similar hand poses. It proposes a new context, different from the one of a single painting where still hands summarise a longer action, and allows researchers to envision these hands under a new perspective. Even though the fundamental meaning of a gesture greatly relies on the iconographic context and the character performing it (Burke, 1992; Dimova, 2020), the idea here is to understand and question the significance of the depicted moments in order to outline specific codes and conventions of representation from the time they were painted. Furthermore, the fact that a hand gesture can or cannot be translated with painted hands not only provides information on the practices of the studied period, but also on the divergences from contemporary ones. Overall, there is a great potential to learn from the data through the browsing process and to develop an understanding of the kind of poses that belong to the narrative spectrum of Early Modern art. However, it is important to acknowledge that the collection used is a subset of paintings from the Italian Early Modern period and results have to be apprehended with this curatorial aspect in mind. Any outcome cannot be considered as a ground truth as it is the result of multiple selections that occurred throughout the shaping and digitization process of the collection. In addition to the curation, the performance of the *OpenPose* model primarily used for hands detection impacts on the results proposed, with an estimated accuracy of 32%. Many hands are thereby missing, whereas others, in spite of manual corrections, have misplaced keypoints, resulting in different pose interpretations from the machine than what they actually represent.

Bibliography

Bell, P., Schlecht, J. and Ommer, B. (2013). Nonverbal Communication in Medieval Illustrations Revisited by Computer Vision and Art History. *Visual Resources Journal, Special Issue on Digital Art History*, **29**. Taylor & Francis: 26–37.

Burke, P. (1992). The language of gesture in early modern Italy. *A Cultural History of Gesture*. Cornell University Press, p. 14.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E. and Sheikh, Y. (2021). OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**(1): 172–86 doi:10.1109/TPAMI.2019.2929257.

Derry, L., Duhaime, D., Kruguer, J., Rodighiero, D., Schnapp, J. and Pietsch, C. (2021). *Surprise Machines*. (Information+ Conference). Online https://vimeo.com/595473865 (accessed 4 February 2022).

Dimova, T. (2020). *Le Langage Des Mains Dans L'art: Histoire, Significations Et Usages Des Chirogrammes Picturaux Aux XVIIe Et XVIIIe Siecles*. 1er édition. Brepols Publishers.

Impett, L. (2020). Analyzing Gesture in Digital Art History. *The Routledge Companion to Digital Humanities and Art History*. Routledge.

Impett, L. and Süsstrunk, S. (2016). Pose and Pathosformel in Aby Warburg's Bilderatlas. *Undefined* /paper/Pose-and-Pathosformel-in-Aby-Warburg%27s-Bilderatlas-Impett-S%C3%BCsstrunk/f3a34525fa7021322f132c80c9517f240cf1e742 (accessed 3 June 2021).

Lenardo, I. di, Seguin, B. L. A. and Kaplan, F. (eds). (2016). *Visual Patterns Discovery in Large Databases of Paintings*.

Seguin, B. (2018). The Replica Project: Building a visual search engine for art historians. *XRDS: Crossroads, The ACM Magazine for Students*, **24**(3): 24–29 doi:10.1145/3186653.

Shen, X., Efros, A. A. and Aubry, M. (2019). Discovering Visual Patterns in Art Collections With Spatially-Consistent Feature Learning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, pp. 9270–79 doi:10.1109/CVPR.2019.00950. https://ieeexplore.ieee.org/document/8954148/ (accessed 18 October 2021).

Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.-L. and Grundmann, M. (2020). MediaPipe Hands: On-device Real-time Hand Tracking. *ArXiv:2006.10214 [Cs]* http://arxiv.org/abs/2006.10214 (accessed 21 December 2021).

Notes

1. For examples see https://www.biblhertz.it/it/photographic-collection and https://www.europeana.eu/fr/collections
2. See projects proposed on the open platform https://artsandculture.google.com/
3. See the ArtLens Wall from the Cleveland Museum of Art https://www.clevelandart.org/artlens-gallery/artlens-wall
4. https://www.biblhertz.it/it/photographic-collection
5. See https://github.com/CMU-Perceptual-Computing-Lab/openpose
6. The library mediapipe was used for the real time extraction of keypoints, see https://google.github.io/mediapipe/(Zhang et al., 2020)

Diachronic Lexicon Induction via Literary Translations

Birkenes, Magnus Breder
magnus.birkenes@nb.no
National Library of Norway

Johnsen, Lars G.
lars.johnsen@nb.no
National Library of Norway

Kåsen, Andre
andre.kasen@nb.no
National Library of Norway

The long term goal of this work is a diachronic or historic lexicon i.e. a mapping of word forms from one time period (loosely defined) to another. A lexicon like this can open new possibilities in information retrieval in historical and cultural heritage collections as well as provide a new foundation for quantitative methods on such material. For example, a modern spelling variant can be used for search and retrieval by utilizing period specific counterparts. The other way around, a collection of words found at different time periods, can be grouped together using a modern variant.

In order to induce the lexicon we make use of a pipeline based on Transformers (Vaswani et al. 2017) for bitext (i.e. corresponding text in at least two languages) mining as in Reimers & Gurevych (2019) and word alignment as in Jalili Sabet et al. (2020). The primary data in this pilot study are two translations of Goethe's *The Sorrows of Young Werther*in Norwegian from 1820 and 1998, respectively, and can be found in the collection of the National Library of Norway.

Just as spoken language changes over time, the written language does as well. Changes in spelling constitutes such a change, sometimes alongside semantic changes. Our main focus lies within spelling variations. Since writing is our only encounter with languages of the not so distant past, so to speak, understanding and mapping this change is important on several levels. One good source for diachronic bitext is collections of an author's complete works which often are modernized from time to time in honor of their birthday and other celebrations. However, such bitexts have

often been modernized in a particular way, according to specific preconceptions or stylistics. But since variation between time periods poses serious challenges to search and retrieval in large collections, and also to quantitative analysis in general, there is a great need for linking such variation. A canonical example is if a word form or spelling variant simply changes one character, we no longer have identical forms across time as in Norwegian where the word *woman*at least have had three forms (*qvinde*, *kvinde*, *kvinne*) from the early 1800s to the present.

Recent developments in Transformer-based text processing have eased the need for the amount of specialized data in machine learning. Reimers & Gurevych (2019) show that a multilingual model is preferable in aligning sentences rather than monolingual models. And Jalili Sabet et al. (2020) provide us with the flexibility to make use of efforts like Kummervold et al. (2021) in aligning words. With such a pipeline, inspired by Shi et al. (2021), we get the aforementioned mapping or lexicon in addition to linguistically interesting word pairs. While the resulting list of word pairs can tell us something about language change as well as language policy, it is also useful to search engines where every user will benefit from being able to access texts from earlier stages without domain knowledge or linguistic expertise. In addition to a historical lexicon, the possibility of making an automatic modernisation or paraphrasing of text, or, put differently, how to make texts more readable, is one of the desired offshoots of this pilot study.

Bibliography

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings. In Findings of the Association for Computational Linguistics: EMNLP 2020, Online. Association for Computational Linguistics.

Per E Kummervold, Javier De la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. 2021. Operationalizing a National Digital Library: The Case for a Norwegian Transformer Model. In Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), Reykjavik, Iceland (Online).

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China. Association for Computational Linguistics.

Haoyue Shi, Luke Zettlemoyer, and Sida I. Wang. 2021. Bilingual Lexicon Induction via Unsupervised Bitext Construction and Word Alignment. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Fifty Shades of Twilight: Computationally Comparing Collocations in Twilight and 50 Shades of Grey

Bordalejo, Barbara
Barbara.bordalejo@uleth.ca
University of Lethbridge

Van Zundert, Joris J.
joris.van.zundert@huygens.knaw.nl
Huygens Institute for the History of the Netherlands – Royal Netherlands Academy of Arts and Sciences

Neugarten, Julia
julia.neugarten@huygens.knaw.nl
Huygens Institute for the History of the Netherlands – Royal Netherlands Academy of Arts and Sciences

We present the results of measuring collocation similarity between *Twilight* (Meyer 2005) and *50 Shades of Grey* (James 2011) . *50 Shades* began as *Twilight*-fanfiction (Brennan and Large 2014). We use these texts for a case study analyzing the transformative effects of fanfiction on the narratives that fans call “canon”. Tosenberger (2014:17) asserts that “fanfiction is given life by what other spaces don’t allow, it […] fills those spaces with stories for which the canon has neither room nor desire.” Fanfiction is a narrative space to explore non-normative topics and perspectives.. *Twilight* narrates the romance of a teenage girl and her vampire boyfriend. *50 Shades* amplifies the mostly unconsummated sexual tension in *Twilight* and eliminates the novel’s supernatural elements. In *50 Shades,* the male protagonist is dangerous not because he is a vampire, but because of his S/M inclinations. Our challenge is to model and quantify these transformations computationally.

Paris (2016) classifies *50 Shades* as "mommy porn" while *Twilight* has been called "abstinence porn" (Seifert 2005). In its vocabulary and collocations, *Twilight* seems like a non-explicit model for *50 Shades*. To test this, we make an educated guess by initially selecting four terms: "soft", "hard", "gaze", and "stare". We hypothesize that words collocating with "soft" and "hard" differ between texts: in *Twilight,* Edward's skin is hard, while in *50 Shades* Christian's penis is hard. Additionally, subjects and objects of stares and gazes differ between texts, with looks conveying love or longing in *Twilight* while conveying sexual desire in *50 Shades*. For each appearance of these terms we compute the pointwise mutual information (PMI) for collocated words in a 9-token context. PMI expresses the probability of a collocation occurring given the occurrence of the individual words (Bouma 2009:3). Window size was based on mean sentence length. A baseline for comparison was computed from the same measure for the YA-novels *Eleanor & Park* (Rowell 2012), *The Fault in Our Stars* (Green 2012), and *Shiver* (Stiefvater 2009). We used Linguistic Inquiry & Word Count (Pennebaker et al. 2015) to calculate the percentage of words related to a specific domain as defined by LIWC's dictionaries within the PMI-results. We then compared the percentage of words associated with the selected terms between the books and compared the LIWC-results for the PMI-data to the LIWC-results for the books as a whole.

Analyzing tokens identified by PMI as significantly collocated with the target words, more words from the LIWC-category "perception" occur around the term "hard" in *50 Shades* than in *Twilight* (9% vs. 7%). *Twilight* shows more perceptions-terms around "soft" (12% vs. 15%). Thus, perceptions are more frequently described as "hard" in *50 Shades* and more frequently as "soft" in *Twilight.* In *Twilight* more verbs occurred around "soft" (20%) than in *50 Shades,* where 13% of significantly collocated tokens for "soft" were verbs. This suggests that more "soft" actions are taken in *Twilight* than in *50 Shades,* which would fit our hypothesis. In *50 Shades,* the word "stare" more frequently occurred near words relating to biological features or processes (9%) than in *Twilight* (3%). Similarly, "gaze" occurred around words relating to biological processes in *50 Shades* (7% *)* and only 5% in *Twilight.* It thus appears that biological processes and parts of the body are often looking and being looked at in both texts, but more frequently in *50 Shades.*

Our analysis confirms that *Twilight* is non-explicit: it scores 0,01% in LIWC's sexuality- and swearing-categories. In LIWC-categories relating to the social and to perception, the texts' score similarly. Our results seem to confirm the hypothesis that *Twilight* can be regarded as the non-explicit counterpart to *50 Shades.* As a next step, we intend to examine the difference in gender-related words in the texts: 4% of words for *Twilight* and 5% for *50 Shades* were male-related, with only 1% female-related words in both . Intuitively this makes sense as both are first-person narratives by female narrators focused on their male love interests . However, male-related words were less frequent in the PMI-results for the selected terms than in the texts as a whole.

Combining PMI and LIWC-results, we developed a method to compare collocations of specific words between texts. This method is a step towards digital hermeneutics, the possibility of "interpreting with digital machines" (Romele, Severo, and Furia 2020:73). During our presentation, we will present more detailed results, baseline comparisons, and will consider possibilities to improve their evaluation and discuss possible next steps such as analysis of word embeddings.

Bibliography

Bouma, G. (2009). Normalized (Pointwise) Mutual Information in Collocation Extraction. *Von Der Form Zur Bedeutung: Texte Automatisch Verarbeiten / From Form to Meaning: Processing Texts Automatically: Proceedings of the Biennial GSCL Conference 2009*. Tübingen: Gunter Narr Verlag, pp. 31–39.

Brennan, J. and Large, D. (2014). 'Let's get a bit of context': 'Fifty shades' and the phenomenon of 'pulling to publish' in 'twilight' fan fiction. *Media International Australia, Incorporating Culture & Policy*(152). Media International Australia, Incorporating Culture & Policy: 27–39 doi: 10.3316/informit.567849615699510. https://search.informit.org/doi/10.3316/informit.567849615699510 (accessed 9 December 2021).

Green, J. (2012). *The Fault in Our Stars*. 1st ed. New York: Durron Books.

James, E. L. (2011). *Fifty Shades of Grey*. 1st ed. Waxahachie: The Writer's Coffee Shop.

Meyer, S. (2005). *Twilight*. London: Atom, Little, Brown Book group.

Paris, L. (2016). Fifty Shades of Fandom: The Intergenerational Permeability of Twilight Fan Culture. *Feminist Media Studies*, **16**(4): 678–92 doi: 10.1080/14680777.2016.1193297. http://www.tandfonline.com/loi/rfms20 (accessed 9 December 2021).

Pennebaker, J. W., Boyd, R. L., Jordan, K. and Blackburn, K. (2015). The development and psychometric properties of LIWC2015 University of Texas at Austin https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf (accessed 9 December 2021).

Romele, A., Severo, M. and Furia, P. (2020). Digital hermeneutics: from interpreting with machines to interpretational machines. *AI & SOCIETY*, **35**(1): 73–86 doi: 10.1007/s00146-018-0856-2. https://doi.org/10.1007/s00146-018-0856-2 (accessed 9 December 2021).

Rowell, R. (2012). *Eleanor & Park*. London: Orion Books.

Seifert, C. (2008). Bite Me! (Or Don't): 'Twilight' Has Created a New YA Genre: Abstinence Porn. *Bitch Media* https://www.bitchmedia.org/article/bite-me-or-dont (accessed 9 December 2021).

Stiefvater, M. (2009). *Shiver*. New York: Scholastic Press.

Tosenberger, C. (2014). Mature Poets Steal: Children's Literature and the Unpublishability of Fanfiction. *Children's Literature Association Quarterly*, **39**(1): 4–27 doi: 10.1353/chq.2014.0010. https://muse.jhu.edu/article/538976 (accessed 9 December 2021).

Keeping Kanji Alive: Lessons in Digital Sustainability

Bosse, Arno

arno.bosse@di.huc.knaw.nl
KNAW Humanities Cluster, The Netherlands

Hibino Lory, Harumi

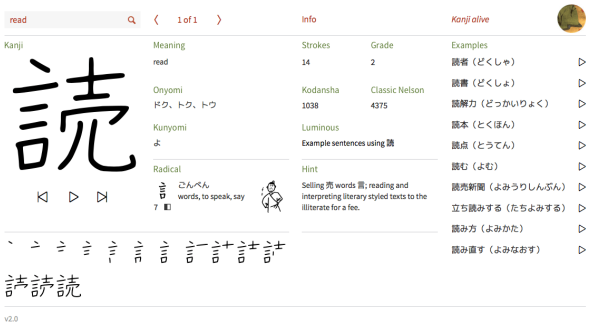
hlory@uchicago.edu
The University of Chicago, USA

Ensuring the sustainability of a digital humanities project in an appropriate and responsible manner remains a critical and vexing challenge for any project team. Part of the reason for this lies in the complex interdependence of the different issues contributing to the challenge—technical, managerial, staffing, financial, and legal—to name but some of the most recognized factors. It is understandable, then, that many practitioners perceive that despite significant progress around digital preservation (for example, on shared digitization or metadata standards) the particular differences between digital projects trying to address long-term sustainability remain far greater than the commonalities.

Funders have recognized this problem as well and (once more, broadly generalising) are providing support to address this problem in different ways. In Europe, for example, where national and EU-wide funding agencies play a significant role, one can observe a renewed commitment to create shared resources for sustainability by investing in large scale digital infrastructures for the humanities (e.g. the EU-funded DARIAH or Clariah in the Netherlands). By contrast, in North-America, given its size, more emphasis has been placed in helping in particular smaller institutions to develop individual strategies for how, and how long, digital projects should be sustained (e.g. in the United-States, the NEH-funded Institute on Advanced Topics on 'Sustaining DH' or in Canada, the SSHRC-funded 'Endings Project').

Both approaches rightly look to address the challenge of sustainability from a future facing perspective. But as a consequence, not enough attention is being paid to the potential lessons we could learn collectively from projects that have already proven themselves as successful and sustainable over an extended period of time. Today, many, if not most digital humanities projects are never conceived to be actively maintained and revised for a significant number of years beyond their initial funding period. The extent of our experience with actually 'sustained' DH projects with active lifespans of five, eight, ten, twelve or more years remains severely limited. Indeed, we tend to regard such projects as 'outliers' whose particular circumstances will likely not overlap with our own. This, in turn, curtails our ability to learn from such projects and examine to what extent their experiences can be applied elsewhere.

In our paper, we would like to present the lessons in sustainability we have learned from continuously maintaining and adapting a digital project over the last twenty years at the University of Chicago. *Kanji alive* (https://kanjialive.com) is a web-application designed to help Japanese language students of all levels learn to read and write kanji. The app is widely used in Japanese language programmes throughout the world with over 70K users/year while the project's main website, which also offers additional language resources for students such as a comprehensive listing of Japanese radicals, has over 150K visitors/year.



Kanji alive is a pedagogical not a research application. However, nearly all of the technical, managerial, staffing, financial, and legal issues central to the sustainability of research driven digital humanities projects apply equally

to successful pedagogical projects as well. Over the past twenty years of its existence, *Kanji alive* has experienced and overcome its fair share of challenges in all of these areas. The purpose and ambition of our short paper is to review and share these lessons with the digital humanities community in the hope that other projects will be able to benefit positively from our experiences as well.

Revealing 'Invisible' Poetry by W. H. Auden through Computer Vision

Using Photometric Stereo to Visualize Indented Impressions in the Poet's Austrian Correspondence and Literary Papers

Brenner, Simon
sbrenner@cvl.tuwien.ac.at
Computer Vision Lab, TU Wien, Austria

Frühwirth, Timo
timo.fruehwirth@oeaw.ac.at
Austrian Centre for Digital Humanities and Cultural Heritage, Austrian Academy of Sciences, Austria

Mayer, Sandra
sandra.mayer@oeaw.ac.at
Austrian Centre for Digital Humanities and Cultural Heritage, Austrian Academy of Sciences, Austria

W. H. Auden (1907-1973) counts among the most influential writers in the English language in the twentieth century. From 1958, the English-born poet divided his time between the United States and Austria: for up to six months of each year, he lived and worked in the Lower Austrian village of Kirchstetten. It is there that, in the period from 1958 to 1973, Auden wrote most of his poetry (Mendelson, 2017: 746, 773; Quinn, 2013: 56; Quinn, 2015: 243). However, while the poet's English and American periods have been extensively researched, W. H. Auden's life and work in Austria are still under-investigated and have attracted scholarly attention only since the 2000s (see, especially, Mendelson, 2005, dedicated to the "European Auden"). It is in the context of an emerging field of Austrian Auden studies (see, especially, Denzer and Seidl, 2014; Neundlinger, 2018) that the Auden Musulin Papers project at the Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH), of the Austrian Academy of Sciences, is situated. The project makes available, through an open-access digital scholarly edition, the letters and literary papers sent by the poet to Welsh-Austrian writer Stella Musulin (1915-1996). These privately owned, and previously inaccessible, documents cast a fresh light on one of Auden's most prolific creative periods.

Some of the documents contain indented impressions of lines of poetry, indicating that Auden reused sheets of paper originally placed underneath those on which he typed his poems. Standard image digitization technologies employed in digital editing, optimized for creating representations of 2D surfaces, cannot capture these three-dimensional impressions. Therefore, the Auden Musulin Papers project takes advantage of computer-vision technologies used in cultural-heritage research of 3D objects. In the context of this project, the Computer Vision Lab (CVL) of TU Wien [Vienna University of Technology] has produced high-resolution photometric-stereo reconstructions of the pages containing typewriter impressions. Photometric stereo (PS) is a computer-vision method that enables the reconstruction of 3D surfaces from a set of images taken under a constant camera view and varying lighting directions. In comparison to other 3D acquisition methods (such as structured light, photogrammetry, or time-of-flight), PS has been demonstrated to be especially efficient for the acquisition of small local surface details (Herbort and Wöhler, 2011; Jackson et al., 2007; McGunnigle and Chantler, 2003; Thumfart et al., 2013). The pages containing the typewriter indentations were imaged with a prototypical PS acquisition system, carrying 54 individually controllable LEDs and an achromatic medium-format camera, originally developed for capturing surface details in medieval parchment manuscripts. The source images acquired of the pages in question have a spatial resolution of 405px/cm (or 1030 dpi), and already in an unprocessed state allow for an improved reading of the indented text. On the basis of surface models generated from those source images, false-color visualizations have been created that clearly distinguish the indented text from other structures in the paper, such as from the texture of the paper itself and overlapping typewriter texts that do contain print ink and leave even more pronounced indentations in the paper. (See https://doi.org/10.5281/zenodo.6458064 for comparing results of standard digital photography, grazing-light source photography, and PS-based false-color visualization.)

The development of such visualizations was completed in early 2022. First tentative results indicate that computer vision affords a unique glimpse into the poet's workshop in Kirchstetten. On the one hand, the 3D reconstructions provide specific evidence of the material writing practices employed by W. H. Auden in the study of his Austrian house. On the other hand, the visualizations of the PS data contribute to our understanding of Auden's poetic practices of composition and revision. Thus, the spatiotemporal relationship between the indented writing, containing

an unpublished version of Auden's poem "Epistle to a Godson," and the handwritten message addressed to Stella Musulin (dated 10 June 1969) over-writing it, provides fresh insights both into the development of the poem and the poet's practices of poetic work.

Bibliography

Denzer, R. and Seidl, M. (eds) (2014). *Silence Turned into Objects: W. H. Auden in Kirchstetten*. St. Pölten: Literaturedition Niederösterreich.

Herbort, S. and Wöhler, C. (2011). An Introduction to Image-Based 3D Surface Reconstructions and a Survey of Photometric Stereo Methods. *3D Research*, 2(3): 1-17.

Jackson, M., Yang, D. and Parkin, R. (2007). Analysis of Wood Surface Waviness with a Two-Image Photometric Stereo Method. *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, 221(8): 1091-99.

McGunnigle, G. and Chantler, M. (2003). Resolving Handwriting from Background Printing Using Photometric Stereo. *Pattern Recognition*, 36(8): 1869-79.

Mendelson, E. (2005). The European Auden. In Smith, S. (ed), *The Cambridge Companion to W. H. Auden*. Cambridge: Cambridge University Press, pp. 55-67.

Mendelson, E. (2017). *Early Auden, Later Auden: A Critical Biography*. Princeton: Princeton University Press.

Neundlinger, H. (ed) (2018). *Thanksgiving für ein Habitat: W. H. Auden in Kirchstetten*. St. Pölten: Literaturedition Niederösterreich.

Quinn, J. (2013). At Home in Italy and Austria, 1948-1973. In Sharpe, T. (ed), *W. H. Auden in Context*. Cambridge: Cambridge University Press, pp. 56-66.

Quinn, J. (2015). Auden's Cold War Fame. In Costello, B. and Galvin, R. (eds), *Auden at Work*. Basingstoke: Palgrave Macmillan, pp. 231-49.

Thumfart, S., Palfinger, W., Stöger, M. and Eitzinger, C. (2013). Accurate Fibre Orientation Measurement for Carbon Fibre Surfaces. In Wilson, R., Hancock, E., Bors, A. and Smith, W. (eds), *Computer Analysis of Images and Patterns: Proceedings of the 15th International Conference on Computer Analysis of Images and Patterns, CAIP 2013*. Springer, pp. 75-82.

Relating the Unread: A Data-Rich Approach to the Literary Canon and the "Great Unread"

Brottrager, Judith

judith.brottrager@tu-darmstadt.de
TU Darmstadt, Germany

Since the emergence of computational approaches in literary studies, the inclusion of more than established and thoroughly researched literary works has been a crucial argument for using quantitative methods despite the inevitable loss of detail caused by necessary formalisations and operationalisations. Franco Moretti (2013: 48 f), for example, argues that so-called *Distant Reading* approaches enable the modelling of an alternative literary history that does not exclude the "Great Unread", i.e. works of literature which have been excluded from the canonical spheres of literary history. Numerous digitisation projects have been launched with this premise in mind: the more texts are available as high-resolution scans, or even better, transcribed as digital text, the broader this alternative literary history can be defined. However, recent contributions to the field (Algee-Hewitt et al., 2016; Underwood and Sellers, 2016; Porter, 2018; Underwood, 2019: 68–110) have shown that besides an ever-expanding archive of digitally available texts, additional data is needed to embed the analysed texts in what Kathrine Bode calls a "data-rich literary history" (2018: 37–57).

This contribution exemplifies how a bespoke context-rich dataset can be compiled by describing the theory-driven creation process of a dataset for the comparative analysis of approximately 1,200 canonised and non-canonised English and German novels and narratives from 1688 to 1914. The description of this use case will focus on the encoding of literary history on two main levels: first, the canon-conscious corpus selection, and second, the data-based operationalisation of canonisation and contemporary reception.

For the corpus compilation, an approach suggested by Algee-Hewitt and McGurl (2015) for the creation of a representative corpus of 20 th century English literature, which focuses on the counterbalancing of inherent biases in available data, has been systematically adapted for the time frame in question. When research projects rely solely on standard collections of already digitised material for their corpus creation, they work with what Algee-Hewitt and McGurl would call a "found" corpus, which builds on layered selection processes that are not transparent but generally linked to a text's status in the canon. To compensate these biases, Algee-Hewitt and McGurl suggest to move from a "found" to a "made" list of texts: By using a predefined list of works to be included in a corpus, gaps in the digitised archive are made visible and can be filled by retro-digitisation (see also Algee-Hewitt et al., 2016).

Even though a canon-conscious corpus selection encodes some aspects of literary history by reconstructing a text's "history of transmission" (Bode, 2018: 38) in terms

of its availability and accessibility, additional data is needed to operationalise literary categories such as canonisation and contemporary reception as numerical scores to be able to use them for quantitative analyses. For both operationalisations, categories suggested by Heydebrand and Winko (1996) have been used. Defining a text's canonisation status as the result of consecutive selective processes, Heydebrand and Winko (1996: 222–23) propose, among others, the continuous scientific engagement with the text (formalised as student editions), interest in its author (formalised as complete/collected works editions), and its treatment in literary history (formalised as mentions in narrative literary histories and other secondary sources) as markers for canonisation. These proxies encompass blurrier features of canonisation, such as longevity (see Bloom, 1994; Assmann, 2008) and cultural capital (see Bourdieu, 1986; Guillory, 1998), while being more generalisable and widely available than publishing records. Analogously to the canonisation status, contemporary reception can be modelled by collecting instances of value judgements, i.e. reviews, from representative journals (*The Monthly Review*, *The Critical Review*, *La Belle Assemblée*, *Flowers of Literature*, *The Star*, *The Athenaeum*, *Allgemeine Literatur-Zeitung*, *Morgenblatt für gebildete Stände*, *Blätter für literarische Unterhaltung, Deutsche Literaturzeitung*) and implicit markers of audiences' interests, as, for example, entries in circulating libraries, which are, in contrast to sales numbers, more representative of lay audience's reading habits (Martino, 1990; Gamer, 2000). Both reviews and circulating library catalogues represent to a certain extent samples of convenience, as the existence of digital surrogates was a prerequisite for their inclusion in the dataset.

Especially for markers of reception and evaluation, the dataset also draws heavily from already existing databases (e.g. British Fiction 1800-1829, English Short Title Catalogue (ESTC), VD17, VD18, Gelehrte Journale und Zeitungen der Aufklärung, The Athenaeum Projects) and is in turn also designed to be sustainable, compatible, and re-usable by adhering to community standards, including international identifiers (as, for example, VIAF), and providing open access and documentation.

Bibliography

Algee-Hewitt, M., Allison, S., Gemma, M., Heuser, R., Moretti, F. and Walser, H. (2016). Canon/Archive. Large-scale Dynamics in the Literary Field. *Pamphlets of the Stanford Literary Lab*(11) https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf.

Algee-Hewitt, M. and McGurl, M. (2015). Between Canon and Corpus: Six Perspectives on 20th-Century Novels. *Pamphlets of the Stanford Literary Lab*(8) https://litlab.stanford.edu/LiteraryLabPamphlet8.pdf.

Assmann, A. (2008). Canon and Archive. In Erll, A., Nünning, A. and Young, S. B. (eds), *Cultural Memory Studies: An International and Interdisciplinary Handbook*. Berlin; New York, NY: Walter de Gruyter, pp. 97–107.

Bloom, H. (1994). *The Western Canon: The Books and School of the Ages*. New York, NY: Harcourt Brace.

Bode, K. (2018). *A World of Fiction: Digital Collections and the Future of Literary History*. Ann Arbor, MI: University of Michigan Press.

Bourdieu, P. (1986). The Forms of Capital. In Richardson, J. (ed), *Handbook of Theory and Research for the Sociology of Education*. Westport, CT: Greenwood, pp. 241–58.

Gamer, M. (2000). *Romanticism and the Gothic: Genre, Reception, and Canon Formation*. Cambridge, UK; New York, NY: Cambridge University Press.

Guillory, J. (1998). *Cultural Capital: The Problem of Literary Canon Formation*. Chicago, IL: University of Chicago Press.

Heydebrand, R. von and Winko, S. (1996). *Einführung in die Wertung von Literatur: Systematik - Geschichte - Legitimation*. Paderborn: Schöningh.

Martino, A. (1990). *Die Deutsche Leihbibliothek: Geschichte Einer Literarischen Institution (1756-1914)*. Wiesbaden: Harassowitz.

Moretti, F. (2013). *Distant Reading*. London, UK; New York, NY: Verso.

Porter, J. D. (2018). Popularity/Prestige. *Pamphlets of the Stanford Literary Lab*(17) https://litlab.stanford.edu/LiteraryLabPamphlet17.pdf.

Underwood, T. (2019). *Distant Horizons: Digital Evidence and Literary Change*. Chicago: The University of Chicago Press.

Underwood, T. and Sellers, J. (2016). The Longue Durée of Literary Prestige. *Modern Language Quarterly*, 77(3): 321–44 doi:10.1215/00267929-3570634.

VisColl 2.0 and VCEditor. A new model and tool in the quiver of codicologists and bibliographers

Campagnolo, Alberto
alberto.campagnolo@gmail.com
Université Catholique de Louvain, Louvain-la-Neuve, Belgium

Porter, Dot

dorp@pobox.upenn.edu
Schoenberg Institute for Manuscript Studies, University of Pennsylvania Libraries

Emery, Doug
emeryr@pobox.upenn.edu
Schoenberg Institute for Manuscript Studies, University of Pennsylvania Libraries

Perkins, Patrick
pperkins@upenn.edu
Schoenberg Institute for Manuscript Studies, University of Pennsylvania Libraries

Ransom, Lynn
lransom@upenn.edu
Schoenberg Institute for Manuscript Studies, University of Pennsylvania Libraries

Introduction

Today's quintessential book form is the codex, i.e., collections of sheets folded double and fastened at the spine, usually protected by covers (Roberts and Skeat, 1983: 1; Ligatus, 2015a; Harnett, 2017: 184). Historically this has been true for most of the book production in the Middle East and the West since the appearance of codex books around the third century CE. The ultimate working unit of books in codex format is the gathering (or quire): a group of folded (or single) leaves bound together with other gatherings to form the textblock (Andrist et al., 2013: 50; Ligatus, 2015b).

The gathering structure represents the first key to studying the genesis and history of codices and their content (Andrist et al., 2010). This structure is fundamental to studying the codex format for manuscripts and early printed books. The study of gathering structures helps assess provenance and dating or illustrate—highlighting irregularities and discontinuities—complex histories.

Traditionally, gathering structures are described in highly formalized alphanumerical formulaic representations —collation formulas—whose information density hinders the immediacy of their interpretation. To tackle this problem and provide a richer and more flexible information model, Porter and Campagnolo have devised VisColl (Collation Visualization), a system for modelling and visualizing the physical collation of books in codex format (Porter et al., 2017a; Porter et al., 2017b). When gathering assemblies are captured through VisColl, taking advantage of the flexibility of the digital medium, they can, in turn, be readily visualized, thus allowing their study and that of related features, fostering a more immediate understanding and interpretation of the data.

VisColl and VCEditor

VisColl's development team is centred at the Schoenberg Institute for Manuscript Studies (SIMS) at the University of Pennsylvania Libraries. Conceived around 2013, VisColl's model and project have gone through a series of ameliorations that resulted in the publishing of version 2.0 and a new graphical interface, VCEditor in October 2021 (Porter et al., 2021; Perkins and Emery, 2021). VCEditor is based on the VisCodex web application, developed by the Old Books New Science Lab at the University of Toronto, under the direction of Professor Alexandra Gillespie (Ung et al., 2021).

The XML model behind VisColl permits users to collect structured data to describe gathering assemblies. From that one can generate diagrammatic representations and digital facsimiles that re-join into virtual bifolia the photographs of the leaves of books, making overt their physical makeup. To allow for maximum flexibility and to represent complex structures, version 2.0 changes the approach and its data elements radically, switching from the quire and the bifolium as its base units (as it is customary in collation formulas) to the single leaf; each leaf can then be conjoined to form a bifolium within a gathering. Users can then collect richer data than what is traditionally captured in collation formulas—e.g., each leaf's attachment method, the hair and flesh side of parchment leaves, animal species, and so on. The data collected can then be used to automatically generate detailed diagrammatic visualizations, recording and describing complex structures with ad hoc annotations and taxonomies. This alone represents a substantial improvement over traditional methods based on collation formulas and descriptions, or diagrams painstakingly produced for a specific volume.

VCEditor, based on UToronto's VisCodex (Ung et al., 2021), allows users to model gathering structures through a simple graphic interface without considering or understanding VisColl's model, thus widening the user base of the model beyond the digital humanities. Data on the structure is collected and mapped onto VisColl's schema to generate compliant XML records. The XML record can then be downloaded or used to generate, directly in VCEditor, a series of visualizations through Idrovora, an HTTP Xproc service developed by Gregor Middell (Figure 1)(Middell, 2021). Idrovora runs XSLT scripts in the backend and generates SVG visualizations of the gatherings, HTML rendering of the reconstructed conjoined bifolia, and collation formulas. The system is very flexible, and additional visualizations or other forms of data reuse can be added through new XLST pipelines or parametrical input.

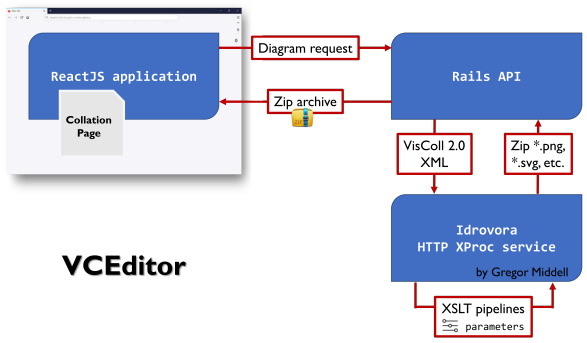


Figure 1.
VCEditor's infrastructure.

Since VCEditor's beta release in early October 2021 to the time of writing, more than 250 people have registered for the service. The users vary from scholars to book conservators and students working on more than 170 individual projects. A testimony to the need and popularity of the tool.

While working on improving the tool and the user experience, we are also actively putting together new scripts for a variety of different visualizations and modelling requirements, such as imposition for printed books and manuscripts penned before folding and cutting the gatherings, watermark positioning, improved biocodicology data visualizations, engaging with users and specific communities of interest.

This presentation will showcase VCEditor's features and the modelling capabilities of VisColl's 2.0.

Bibliography

Andrist, P., Canart, P. and Maniaci, M. (2013). *La syntaxe du codex: essai de codicologie structurale*. (Bibliologia 34). Turnhout: Brepols.

Harnett, B. (2017). The Diffusion of the Codex. *Classical Antiquity*, **36**(2): 183–235 doi:10.1525/ca.2017.36.2.183.

Ligatus (2015a). Codex-form Books *Language of Bindings*. London: University of the Arts London https://www.ligatus.org.uk/lob/concept/4886 (accessed 27 December 2020).

Ligatus (2015b). Gatherings *Language of Bindings*. London: University of the Arts London http://w3id.org/lob/concept/2286 (accessed 21 March 2021).

Middell, G. (2021). *Idrovora: A Pump Station for Your XProc Pipelines*. Clojure https://github.com/gremid/idrovora (accessed 10 April 2022).

Perkins, P. and Emery, D. (2021). *VCEditor*. JavaScript Philadelphia (PA): University of Pennsylvania Libraries https://github.com/KislakCenter/VisualCollation (accessed 3 December 2021).

Porter, D., Campagnolo, A. and Connelly, E. (2017a). VisColl: A New Collation Tool for Manuscript Studies. In Busch, H., Fischer, F. and Sahle, P. (eds), *Kodikologie & Paläographie Im Digitalen Zeitalter 4 | Codicology & Palaeography in the Digital Age 4*. Norderstedt: Books on Demand Gmbh, pp. 81–100 https://kups.ub.uni-koeln.de/7782/.

Porter, D., Campagnolo, A., Tuohy, C. and Emery, D. (2021). *VisColl: Release v2.1.0*. Philadelphia (PA): University of Pennsylvania Libraries doi:10.5281/zenodo.5139928. https://doi.org/10.5281/zenodo.4075530 (accessed 3 December 2021).

Porter, D., Gillespie, A., Campagnolo, A., Mitchell, L. and Di Cresce, R. (2017b). VisColl: Visualizing the physical structure of medieval manuscripts, a poster and demonstration. *Digital Humanities 2017*. Montréal: McGill University; Université de Montréal, pp. 778–79 https://dh2017.adho.org/abstracts/DH2017-abstracts.pdf.

Roberts, C. H. and Skeat, T. C. (1983). *The Birth of the Codex*. London; New York: Published for the British Academy by the Oxford University Press.

Ung, M., Rajakumar, J., Asghar, I., Law, D. and Liu, S. (2021). *VisCodex*. JavaScript Toronto: University of Toronto Libraries https://github.com/utlib/VisualCollation (accessed 29 August 2019).

Nursing the Subaltern: Using Digital Prosopography to Explore the Transnational Makings of Filipino Nurses Since 1898

Capucao. Jr., Reynaldo Caasi
rcc9vq@virginia.edu
University of Virginia, United States of America

Historical scholarship in nursing has proven significant in elucidating gender roles in medicine and society but often disregards race and ethnicity in its framing of the middle-class white woman as the focus of its discourse. One ethnic group at the periphery of nursing history are Filipinos, as evident with the historiography failing to mirror these nurses' present-day hypervisibility within the global healthcare arena. This relationship between the Philippines and professional nursing, however, can

be traced to the importation of the United States (US) model of nursing following the aftermath of the Spanish American War. Although the Philippines gained autonomy in 1946, its Americanized nurses still don the traditional white uniform and continue to migrate in droves to placate US labor shortages. The ahistorical Filipino nurse subject, through ongoing migrations, haunts the legacy of American colonialism and its imperialist agenda of creating an overseas market in Asia.

My attempt to capture this elusive subject along its ghostly passage occurs through my use of digital prosopography. This historical method has the capacity to collectively capture the ephemeral identity of the Filipino nurse. [i] To address the limits of the Filipino nurse archive, I focus my study on autobiography instead of biography as a source of life history. As Peter Cunningham demonstrates the potential for oral history in prosopographical study, I draw upon oral history interviews of Filipino nurses located in the U.S. for this pilot study. My corpus currently contains thirteen interviews: four archived interviews from Seattle, Washington and nine interviews from Virginia conducted by me. The structure of interviews varies due to intersubjectivity and dissimilar interviewers and project objectives among them, but the aggregate covers the lives of subjects.

Despite the small corpus, digital humanities approaches have the capacity to visualize data and create new modes of interpretation from a few records–a problem that persists among subaltern archives. Approaches used to extend the limits of my archive entail mid-range reading with markup, constructing a searchable database, and georeferencing. My use of mid-range reading adapts the Biographical Elements and Structural Schema (BESS) tagging method implemented by the *Collective Biographies of Women* (*CBW*) to explore narrative structure and composition of responses. BESS explores particularities of national identity, occupation, and space, and therefore, provides a model for my analysis of oral history interviews. Instead of an analysis at the paragraph-level, I segment and number portions of text in correspondence to discourses around geography and the built environment. Tagging involves categories of persona type, events, persona description, discourse, and topos. This process begins by digitizing interview transcripts to undergo markup via Oxygen XML editor, and the end goal is to produce a searchable database of categories that is linked to the *CBW* website. Tags thus provide the basis for distant scale analysis to establish relational networks and typologies.

Events (observable actions in time and space) tagged can be further defined with attributes to a particular event, which includes georeferencing. For example, events tagged as, “travel for work,” can include attributes for specific and non-specific location structures or settings and dates. The array of attributes reflects initial findings of highly mobile, professional female immigrants. The immigration dates of subjects to the U.S. range between 1926 and 2004. Although subjects settled in Seattle or Virginia, these locations were not usually the sites of arrival. The number of migrations prior to their final destinations vary greatly and must be contextualized spatially and historically. Mapping, thus, as a digital tool, allows visual comparisons among groups within my corpus: place (Seattle and Virginia) and immigration dates (1925-1950, 1951-1975, and 1976-). Beyond the mapping of toponyms, I highlight the capability of mapping as a critical practice using ArcGIS Pro to highlight the significance of prosopographical findings in delineating upon human relationships, migration, and the built environment.

This project illuminates the individual voices of Filipino nurses which serve as a critique to view the contradictions of American liberal ideals. But the scarcity of oral history interviews, particularly of Filipino nurses who migrated to the U.S. during the first half of the twentieth century, limits the ability of systematic investigation. The capacity of digital humanities approaches, however, can make meaningful analyses from a limited archive to then limn a collective portrait of the Filipino nurse identity. Although the project is in its preliminary stages, it has begun the process of excavating this identity through a myriad of digital methods and tools.

Bibliography

Booth, A. (2017). Mid-range reading: Not a manifesto. *PMLA/Publications of the Modern Language Association of America*, **132**(3): 620-627, DOI: 10.1632/pmla.2017.132.3.620.

Bullough, V.L., Bullough, B., and Wu, Y.B. (1992). Achievement of eminent American nurses of the past: A prosopographical study. *Nursing Research*, **41**(2): 120-124.

Choy, C. C. (2003). *Empire of Care: Nursing and Migration in Filipino American History*. Durham: Duke University Press.

Cunningham, P. (2001). Innovators, networks and structures: Towards a prosopography of progressivism. *History of Education*, **30**(5): 433-451, DOI: 10.1080/00467600110064726.

Gordon, Avery (2008). *Ghostly Matters: Haunting and the Sociological Imagination*. Minneapolis: University of Minnesota Press.

Hawkins, S. (2012). *Nursing and Women's Labour in the Nineteenth Century: The Quest for Independence*. Abingdon, Oxfordshire, UK: Routledge.

Lowe, L. (1996). *Immigration Acts: On Asian American Cultural Politics*. Durham: Duke University Press.

Pollitt, P. and Humphries, A. (2013). Nursing in a time and place of peril: Five North Carolina nurses. *Journal of Nursing Education and Practice*, **3**(9): 176-186. DOI: 10.5430/jnep.v3n9p176.

Spires, K. A. (2013). Nurses in the Boer War (1899-1902): What was it about the collective body of nurses caring for the sick and wounded during the Boer War that shaped the future of military nursing? Ph.D. thesis, London South Bank University.

Sy, L. (2020). BESS: A very short primer. *Scholar's Lab Blog*, https://scholarslab.lib.virginia.edu/blog/bess-primer/ (accessed 8 December 2021).

Verboven, K., Carlier, M., and Dumolyn J. (2007). A short manual to the art of prosopography. In Keats-Rohan, K. (ed),*Prosopography Approaches and Applications: A Handbook*. Oxford: University of Oxford, pp. 35-70.

Notes

i. Prosopographical research in nursing remains an underutilized method, but when used, it examines the history of Western civilization.

Take a sip of TEI and relax: a proposition for an end-to-end workflow to enrich and publish data created with automatic text recognition

Chagué, Alix
alix.chague@inria.fr
Inria, France; Université de Montréal, Canada

Scheithauer, Hugo
hugo.scheithauer@inria.fr
Inria, France

Terriel, Lucas
lucas.terriel@inria.fr
Inria, France

Chiffoleau, Floriane
floriane.chiffoleau@inria.fr
Inria, France

Tadjo Takianpi, Yves
yves.tadjo-takianpi@inria.fr
Inria, France

Romary, Laurent
laurent.romary@inria.fr
Inria, France

Over the last decades, several breakthroughs have made the dream to automatically transcribe thousands of handwritten documents a reality (Causer et al., 2018; Sánchez et al., 2017; Seaward, 2017; Yin et al., 2013). For example, software like *Transkribus* (Kahle et al., 2017) and *eScriptorium* (Stokes et al., 2021) provide non-specialist users with simple environments to conduct transcription campaigns relying on efficient HTR [1] engines. While transposing scriptures from a piece of paper onto a text editor used to require effort and concentration, it is now possible to imagine simply pressing a button and letting your computer work while you start preparing your next cup of tea. A few minutes later, your drink is ready, and so is the transcription of the two thousand pages you needed. As automatic transcription software is about to produce huge volumes of data (Clanuwat et al., 2019; Camps, 2021. See also the Vietnamica project [2].), it seems crucial to think about how we can interact with the resulting files with maximum efficiency.

In response to previous similar initiatives (Carius, 2020), we would like to present an end-to-end workflow revolving around the use of various automatic techniques to go from a set of digital images to the actual publication of a text edition. Such techniques include, on top of HTR, information extraction tools [3] and an open source and ready-to-use environment for publication. Moreover, we aim to make this framework as simple and generic as possible: it is independent from the transcription engine, and potentially compatible with any language, writing system, and any type of document (Balogh and Griffiths, 2020. See also the TEI Special Interest Group for East Asian/Japanese [4]).

Several key principles ensure the coherence of the workflow: transparency and availability of the data at each step and the use of a fully standardized format like TEI XML as the cornerstone to store all the available information. Other XML standards like ALTO [5] or PAGE (Pletschacher & Antonacopoulos, 2010) are commonly used by transcription software to export the output, but we advocate for a change of paradigm in order to give more importance to TEI earlier in the workflow (Scheithauer et al., 2020). The TEI guidelines define a set of elements to document this type of data, namely "sourceDoc" and its children [6]. Leveraging TEI from the start is essential to connect the metadata of the images [7] and documents, the text and layout information generated during the

transcription, and any further editorial layer added to the raw transcription.

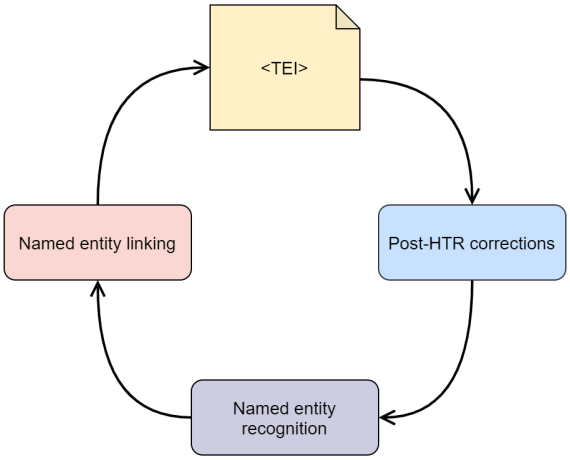


Fig. 1:
TEI as a threefold structure.

We imagine a configuration capable of processing a large family of TEI customizations as long as the file follows a structure (Fig. 1) in which:

- "teiHeader" stores the metadata,
- "sourceDoc" the raw transcription, and
- "body" the interpreted logical structure along with the editorial layers [8].

We thus aggregate two phases in the digitization lifecycle which are often disconnected.

Editorial operations can include preprocessing tasks such as post-HTR corrections (spell-checking) and text normalization, as well as information extraction (text mining). When the volume of data increases, extracting and linking named entities with indexes quickly risks becoming a laborious task. Instead, natural language processing tools can automate the process (Ehrmann et al., 2020; Frontini et al., 2015) all the while relying on the analysis of the sentences and words within their context. We developed *Semantic@*, a proof of concept utilizing deep learning models, to extract named entities which are then cycled back into the TEI tree (Fig. 2). The extraction of named entities (i.e. names of people, places, or dates, etc.) is a crucial step before disambiguation which further permits to build links with open general or domain-specific knowledge bases. These steps allow for later explorations of the text with data mining technologies.

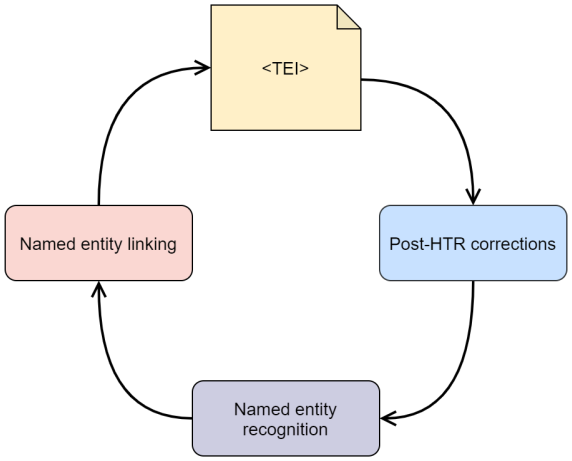


Fig. 2:
Virtuous circle for the enriched TEI document.

Once all the layers of an edition are connected into the same TEI file, edited documents can be posted online with softwares like *TEI Publisher* (Turska et al., 2016; Chiffoleau et al., 2021). It provides a fully customizable environment where templates generate "views" based on the content of the XML files. With the aforementioned TEI structure, we propose an edition template containing:

1. a flat representation of the transcription,
2. an imitative representation of the transcription based on SVG [9] integrating the layout of the pages,
3. a diplomatic edition of the source document, based on the content of the body element, and
4. a facsimile, using the IIIF protocol (Fig. 3).

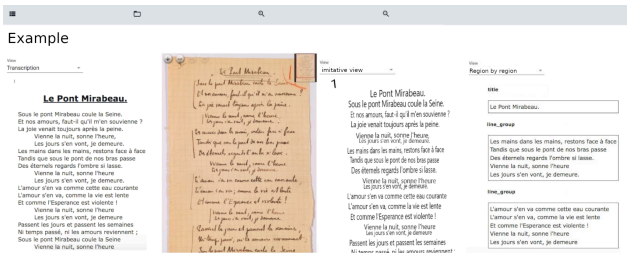


Fig. 3:
A mock-up showing the four different views potentially available in an application like TEI-Publisher.

We would like to take the opportunity of presenting a short paper during the DH2022 international conference to subject our framework (Fig. 4) -and its robustness to different writing systems- to the scrutiny of the DH community. In particular, we believe that our proposition addresses challenges raised by Open Science, primarily the necessity to gain better control over every step

within complex pipelines that involve various tools, thus facilitating reproducibility. A paradigm revolving around a pivotal element, like a TEI file grouping the different results, frees us from the constraint of a linear progression by maintaining multiple entry points in the workflow.

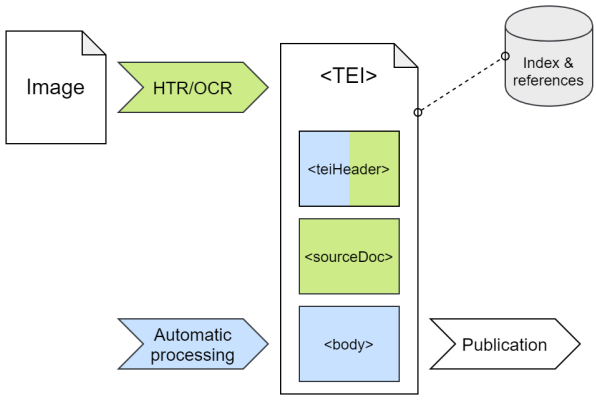


Fig. 4:
Simplifying the workflow by using TEI from the beginning.

Bibliography

Balogh, D. and Griffiths, A. (2020). DHARMA Encoding Guide for Diplomatic Editions EFEO ; Humboldt-Universität (Berlin) ; CEAIS - Centre d'Études de l'Inde et de l'Asie du Sud report https://halshs.archives-ouvertes.fr/halshs-02888186 (accessed 10 December 2021).

Camps, J.-B. (2021). Gallic(orpor)a: Extraction, annotation et diffusion de l'information textuelle et visuelle en diachronie longue Paper presented at the Inauguration du BnF DataLab, Paris https://www.academia.edu/58990010/Gallic_orpor_a_Extraction_annotation_et_diffusion_de_l_information_textuelle_et_visuelle_en_diachronie_longue (accessed 9 December 2021).

Carius, J.-C. (2020). Plateforme d'éditions enrichies à l'INHA : Premier point d'étape d'un projet en cours d'élaboration Billet *Numérique et recherche en histoire de l'art* https://numrha.hypotheses.org/1107 (accessed 8 December 2021).

Causer, T., Grint, K., Sichani, A.-M. and Terras, M. (2018). 'Making such bargain'': Transcribe Bentham and the quality and cost-effectiveness of crowdsourced transcription. *Digital Scholarship in the Humanities* doi: 10.1093/llc/fqx064. https://academic.oup.com/dsh/advance-article/doi/10.1093/llc/fqx064/4810663 (accessed 11 June 2018).

Chagué, A. and Scheithauer, H. (2021). *LEPIDEMO, a Pipeline Demonstrator for LECTAUREP to Go from EScriptorium to TEI-Publisher*. Jupyter Notebook doi: 10.5072/zenodo.977657. https://github.com/lectaurep/lepidemo (accessed 10 December 2021).

Chiffoleau, F., Baillot, A. and Ovide, M. (2021). A TEI-based publication pipeline for historical egodocuments -the DAHN project. *Next Gen TEI, 2021 - TEI Conference and Members' Meeting*. Virtual, United States https://hal.archives-ouvertes.fr/hal-03451421 (accessed 10 December 2021).

Clanuwat, T., Lamb, A. and Kitamoto, A. (2019). KuroNet: Pre-Modern Japanese Kuzushiji Character Recognition with Deep Learning. *ArXiv:1910.09433 [Cs]* http://arxiv.org/abs/1910.09433 (accessed 8 December 2021).

e-editiones (2021). *Eeditiones/Tei-Publisher-App*. XQuery e-editiones.org https://github.com/eeditiones/tei-publisher-app (accessed 10 December 2021).

Ehrmann, M., Romanello, M., Flückiger, A. and Clematide, S. (2020). Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers. *CLEF 2020 Working Notes. Conference and Labs of the Evaluation Forum*. [online event]: Zenodo doi: 10.5281/ZENODO.4117566. https://zenodo.org/record/4117566 (accessed 10 December 2021).

Frontini, F., Brando, C. and Ganascia, J.-G. (2015). Semantic Web Based Named Entity Linking for Digital Humanities and Heritage Texts. In Zucker, A., Draelants, I., Zucker, C. F. and Monnin, A. (eds), *First International Workshop Semantic Web for Scientific Heritage at the 12th ESWC 2015 Conference*. Portorož, Slovenia: Arnaud Zucker and Isabelle Draelants and Catherine Faron Zucker and Alexandre Monnin https://hal.archives-ouvertes.fr/hal-01203358 (accessed 10 December 2021).

Kahle, P., Colutto, S., Hackl, G. and Mühlberger, G. (2017). Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 04. pp. 19–24 doi: 10.1109/ICDAR.2017.307.

Kiessling, B. (2021). *Mittagessen/Kraken*. Python https://github.com/mittagessen/kraken (accessed 10 December 2021).

Lopez, P. (2008). *GROBID*. Java https://github.com/kermitt2/grobid (accessed 10 December 2021).

Pletschacher, S. and Antonacopoulos, A. (2010). The PAGE (Page Analysis and Ground-truth Elements) format framework. pp. 257–60 doi: 10.1109/ICPR.2010.72.

Sánchez, J. A., Romero, V., Toselli, A. H., Villegas, M. and Vidal, E. (2017). ICDAR2017 Competition on Handwritten Text Recognition on the READ Dataset. IEEE Computer Society, pp. 1383–88 doi: 10.1109/ICDAR.2017.226. https://www.computer.org/csdl/proceedings-article/icdar/2017/3586b383/12OmNy4IEXJ (accessed 9 December 2021).

Scheithauer, H., Chagué, A., Gabay, S., Romary, L., Janes, J. and Jahan, C. (2021). From page to content – which TEI representation for HTR output?. *Next Gen TEI, 2021 - TEI Conference and Members' Meeting*. Weaton (virtual), United States https://hal.archives-ouvertes.fr/hal-03380807 (accessed 7 December 2021).

Seaward, L. (2017). Project Update – teaching a computer to READ Bentham *UCL Transcribe Bentham* http://blogs.ucl.ac.uk/transcribe-bentham/2017/06/09/project-update-teaching-a-computer-to-read-bentham/ (accessed 4 June 2018).

Stern, R. (2013). Identification automatique d'entités pour l'enrichissement de contenus textuels Université Paris-Diderot - Paris VII phdthesis https://tel.archives-ouvertes.fr/tel-00939420 (accessed 10 December 2021).

Stokes, P. A., Kiessling, B., Ezra, D. S. B., Tissot, R. and Gargem, E. H. (2021a). The eScriptorium VRE for Manuscript Cultures. *Classics@ Journal*. [online] https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/ (accessed 30 November 2021).

Stokes, P. A., Kiessling, B., Ezra, D. S. B., Tissot, R. and Gargem, E. H. (2021b). The eScriptorium VRE for Manuscript Cultures. *Classics@ Journal*. [online] https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/ (accessed 30 November 2021).

Terriel, L. (2021). *Semantic@*. Python https://github.com/Lucaterre/semanticat (accessed 10 December 2021).

Tissot, R. (2021). *Scripta/EScriptorium*. Python https://gitlab.com/scripta/escriptorium/-/tree/v0.10.2a (accessed 10 December 2021).

Turska, M., Cummings, J. and Rahtz, S. (2016). Challenging the Myth of Presentation in Digital Editions. *Journal of the Text Encoding Initiative*(Issue 9). Text Encoding Initiative Consortium doi: 10.4000/jtei.1453. https://journals.openedition.org/jtei/1453 (accessed 10 December 2021).

Yin, F., Wang, Q.-F., Zhang, X.-Y. and Liu, C.-L. (2013). ICDAR 2013 Chinese Handwriting Recognition Competition. IEEE Computer Society, pp. 1464–70 doi: 10.1109/ICDAR.2013.218. https://www.computer.org/csdl/proceedings-article/icdar/2013/06628856/12OmNxEBzcq (accessed 9 December 2021).

https://gitlab.com/scripta/escriptorium/-/tree/v0.10.2a

Notes

1. HTR stands for Handwritten Text Recognition.
2. Vietnamica is a research project undertaken jointly by the École Pratique des Hautes Études, the Institute of Hán-Nôm Studies, the Social Sciences Academy of Viêt Nam and the National University of Viêt Nam (Faculty of Humanities and Social Sciences). See https://vietnamica.online/
3. Rosa Stern defined information extraction as a task consisting of extracting and structuring, in semantic classes, the specific information elements contained in non-structured data for automatic processing, such as coreference resolution, relationship extraction, and named entity recognition (Stern, 2013, p. 59).
4. See https://tei-c.org/Activities/SIG/EastAsian/ and https://wiki.tei-c.org/index.php/SIG:East_Asian
5. See the Analyzed Layout and Text Object (ALTO) 4.2 schema specifications at https://www.loc.gov/standards/alto/news.html#4-2-released
6. See https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-sourceDoc.html
7. Including when the images are distributed within the IIIF framework .
8. Logical structure reconstruction can be performed semi-automatically (see the pipeline built for the LECTAUREP project called "LEPIDEMO", https://github.com/lectaurep/lepidemo), or automatically with tools such as GROBID (https://github.com/kermitt2/grobid).
9. An XML-based markup language, see the Scalable Vector Graphics (SVG) 2 recommandations at https://www.w3.org/TR/SVG2/ ; we wish to point at the fact that working with SVG when displaying transcriptions allows us to deal with different writing systems and languages.

The Great Transformation of the Clan System in Early China: A Social Network Analysis of Clan-sign Inscriptions from 1300 BC to 900 BC

Chen, Yuqi
cyq0722@pku.edu.cn
Department of History, Peking University

Wang, Linxu
wanglinxu@pku.edu.cn
Department of Information Management, Peking University

AbstractThis paper analyzes 1206 inscribed bronzes excavated from different archaeological sites. Treating the phenomenon that bronzes with different clan signs

appeared in the same tomb as a kind of co-occurrence, the co-occurrence networks of clan signs in two sequential time periods (1300 BC – 1046 BC and 1046 BC – 900 BC) are constructed to examine the interrelationship of the clans. The strong heterogeneity of the networks in different periods shows the huge impact of the Zhou conquest of Shang, which is called "the Great Transformation from Shang to Zhou" by historians and archaeologists.

Keyword Chinese bronze inscription; clan sign; co-occurrence; historical network analysis

1 BACKGROUND

One well-recognized cultural trait of ancient China was the existence of the clan system, which can be traced to the Shang dynasty and has deeply affected Chinese history for three thousand years.

Due to a lack of historical documents, bronze inscriptions are of great value to the study of clans in early China. During the late Shang dynasty (1300 BC – 1046 BC), the most common form of bronze inscriptions was a single graph functioning as a kind of emblem designating the clan name of the ancestor to whom the bronze was dedicated. After the Zhou people conquered the Shang dynasty, this cultural custom still lasted for a period of time among the Shang people in the early Western Zhou dynasty (1046 BC – 900 BC).

According to our current estimate, there are about 6800 bronzes with the so-called clan signs. 1200 of them are unearthed, while the rest are unprovenanced. A thorough investigation of these materials can increase our understanding of the clan system and fill the gaps in the early history of China.

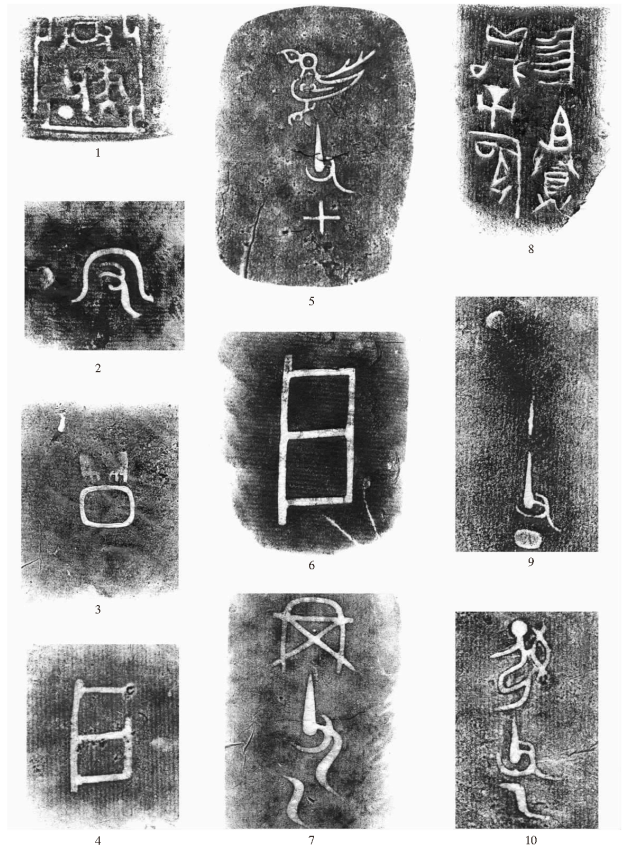




Figure 1.
Bronzes with different clan signs unearthed from Shigushan(石鼓山) M1

2 RELATED WORK

Scholars have been studying the clan-sign inscriptions since the Song dynasty. However, traditional humanities researchers mostly focused on the paleographical study of clan signs or the historical origin of each clan. The inter-clan associations haven't been paid enough attention.

Archaeological discoveries in recent decades provided new clues. Bronze objects belonging to one clan were often found to be buried in tombs of another clan, which can be attributed to gift-giving, marriage, exchange or pillage. Bronzes with different clan signs excavated from the same tomb can reflect the interrelationships between the tomb occupant's clan and other clans.

Some scholars have realized the potential of such investigation (Barnard, 1986; Sun, 2017). But their attempts were all limited to case studies, not from a macro perspective. Using methods of Social Network Analysis, we can discuss the clan system and social structure of early China from a whole new perspective.

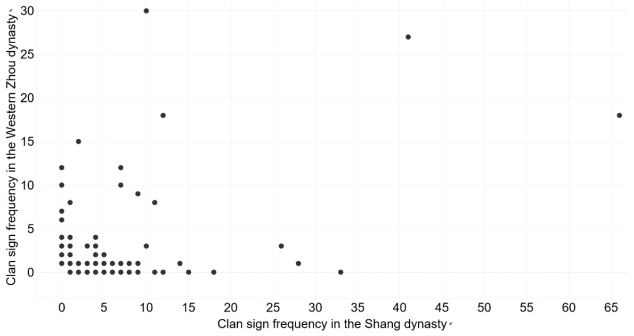
3 DATA AND METHOD

Our data were extracted from *A Compendium of Inscriptions and Images on Bronzes from the Shang and Zhou Period* (商周青銅器銘文暨圖像集成) by manual, including 1206 unearthed bronzes with clan signs. Among them, 684 bronzes were excavated from 219 tombs of the late Shang dynasty, while 522 bronzes were excavated from 185 tombs or hoards of the Western Zhou dynasty. The collected data underwent text preprocessing including tokenization, standardization and manual proofreading. 309 clan signs were recognized on these bronzes.

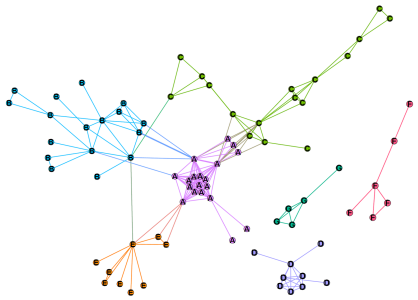
The following methods were then applied:

(1) Frequency analysis of clan signs from the two periods was conducted to show the rise or decline of different clans, since the bronzes were thought to be the symbol of wealth and status.

(2) Treating the archaeological phenomenon that different clan signs appeared in the same tomb as a kind of co-occurrence, the co-occurrence networks of clan signs in the two periods (1300 BC – 1046 BC and 1046 BC – 900 BC) were constructed to examine the interrelationships of these clans.



Clan sign frequency in the Shang and Western Zhou dynasty



Co-occurrence network of clan signs in the late Shang dynasty (1300 BC – 1046 BC)

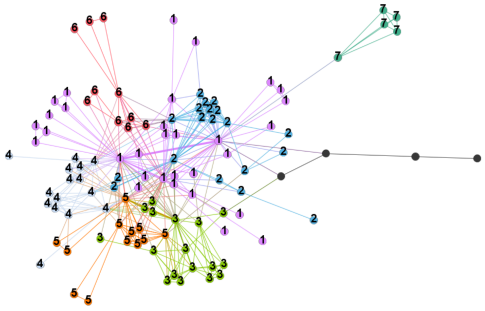


Figure 4.
Co-occurrence network of clan signs in the early Zhou dynasty (1046 BC – 900 BC)

Groups	Clan signs
A	
B	
C	
D	
E	
F	
G	

Table 1.
Clan groups in the late Shang dynasty (1300 BC – 1046 BC)

Groups	Clan signs
1	
2	
3	
4	
5	
6	
7	

Table 2.

Clan groups in the early Zhou dynasty (1046 BC – 900 BC)

4 RESULTS AND DISCUSSION

4.1 Changes of the networks over time

The network in the early Western Zhou dynasty (1046 – 900 BC) is denser than the network in the late Shang dynasty (1300 – 1046 BC), although the size of nodes in the two networks are almost the same. More frequent co-occurrences of clan signs reflect more frequent associations between clans.

Table 3. Metrics of the co-occurrence networks of clan signs

Period	Clan signs	Edges	Density	Average degree	Average weighted degree	Average path length	Modularity
1300 – 1046 BC	137	261	0.028	3.810	7.737	3.079	0.708
1046 – 900 BC	130	428	0.051	6.585	13.631	2.657	0.526

4.2 Clan groups in different periods

Using the Louvain algorithm with a resolution at 1.0, we partitioned the two networks separately into different communities, which can be seen as the clan groups. We laid out the network with the Yifan Hu algorithm and used different colors for different groups (Figure 3 and 4). Clans of the same group had closer connections.

Besides, we used the Kulczynski distance (Cha, 2007; Shang, 2021) to measure the similarity between groups in different periods. There is no similarity score higher than 0.2, proving that the heterogeneity of the two networks is very strong.

Table 4. Kulczynski similarity between the clan groups in different periods

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7
Group A	0.042	0.091	0.091	0	0.072	0.099	0
Group B	0.053	0.111	0.111	0	0	0	0.126
Group C	0.056	0	0	0.085	0.170	0	0
Group D	0.060	0.063	0	0	0	0	0.133
Group E	0	0.142	0.071	0	0	0.125	0
Group F	0	0.159	0	0	0	0	0
Group G	0	0	0	0	0	0	0

Cultural and political changes from the late Shang to the early Western Zhou dynasty is called “the Great Transformation from Shang to Zhou” by some scholars (Wang, 1923). With methods of network analysis, we examined this traditional topic from a new angle and displayed it in a visual way. The results showed us how deeply the Zhou conquest of Shang changed the inter-clan associations and the entire clan system of the Shang people.

For future work, we plan to build a hyper-network with three different types of nodes (tombs, bronzes and clan-sign inscriptions) to reveal more historical details, which can not

only increase our understanding of the clan system but also contribute to the study of burial custom in early China.

Bibliography

Bastian, M., Heymann, S. and Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proceedings of the International AAAI Conference on Web and Social Media*, 3(1): 361-62.

Barnard, N. (1986). A New Approach to the Study of Clan-sign Inscriptions of Shang. In Chang, K. C. (ed), *Studies of Shang Archaeology*. New Haven: Yale University Press, pp. 141-206.

Cha, S.-H. (2007). Comprehensive Survey on Distance/Similarity Measures Between Probability Density Functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 1(4): 300-307.

Fei, J. C. H. and Liu, T.-J. (1982). The Growth and Decline of Chinese Family Clans. *The Journal of Interdisciplinary History*, 12(3): 375-408.

He, J. (2009). *Studies of Clan Inscriptions on Bronzes in the Shang and Zhou Dynasty*. Jinan: Qilu shushe.

Liu, J., Wang, Z. and Xin, Y. (2013). The Excavation of the Western Zhou Tombs at Shigushan in Baoji, Shaanxi. *Wenwu*, 2: 4-54.

Painter, D. T., Daniels, B. C. and Jost, J. (2019). Network Analysis for the Digital Humanities: Principles, Problems, Extensions. *Isis*, 110(3): 538-54.

Rawson, J. (1985). Late Western Zhou: A Break in the Shang Bronze Tradition. *Early China*, 11: 289-96.

Sun, Y. (2017). Inscribed Bronzes, Gift-giving and Social Networks in the Early Western Zhou: A Case Study of the Yan Cemetery at Liulihe. In Shaughnessy, E. C. (ed), *Imprints of Kinship: Studies of Recently Discovered Bronze Inscriptions from Ancient China*. Hong Kong: The Chinese University Press, pp. 47-70.

Shang, W. (2021). The Aristocratic Social Network in the Eastern Jin Dynasty (317-420 C.E.) [Panel discussion]. Historical Network Research in Chinese Studies. (Online conference)

Wang, G. (1923). Institutional Change in the Yin and Zhou Dynasties. In Wang, G., *Guantang Jilin*. Shanghai: Wucheng Jiangshi kanben.

Wu, Z. (2012). *A Compendium of Inscriptions and Images on Bronzes from the Shang and Zhou Period*. Shanghai: Shanghai guji chubanshe.

Machines Reading Maps: from text on maps to linked spatial data

Chiang, Yao-Yi
yaoyi@umn.edu
University of Minnesota

Holmes-Wong, Deborah
dhwong@usc.edu
University of Southern California

Kim, Jina
kim01479@umn.edu
University of Minnesota

Li, Zekun
li002666@umn.edu
University of Minnesota

McDonough, Katherine
kmcdonough@turing.ac.uk
The Alan Turing Institute

Simon, Rainer
rainer.simon@ait.ac.at
Austrian Institute of Technology

Vitale, Valeria
vvitale@turing.ac.uk
The Alan Turing Institute

Maps constitute a significant body of global cultural heritage, and the number of digitized maps is only growing. However, the lack of metadata makes the right maps hard to find: the content of many collections therefore remains opaque to researchers and the general public alike. In this paper, we discuss a digital workflow to create machine-readable data from text on maps, both as a means to make cartographic collections more accessible and interconnected, and as a source of unique historical, geographical and anthropological information.

Usually, critical investigation of maps continues on a small scale, through the close 'reading' of a few items. Digitisation has brought attention to place names featured on historical maps as well as other textual labels that represent a key source for analyzing 'platial'

knowledge. But how can we create large-scale datasets and systematically explore that information?

Projects like *GB1900* or the National Library of Scotland's Map Transcription Projects use volunteers to transcribe words printed on maps (Aucott & Southall 2019). Such efforts are resource intensive and hardly scalable. At the same time, the graphic style of historical maps presents a number of challenges that have hindered automatic recognition of text on maps. *Machines Reading Maps* (MRM) has been working to address these issues by improving and extending existing technologies, and applying standards and best practice to make our outputs FAIR (findable, accessible, interoperable, reusable).

Setting aside the artificial opposition between manual and automatic annotation as mutually exclusive modes of working with maps, MRM's workflow explores what can be gained from their interaction. We integrate a custom version of the annotation platform *Recogito* (Vitale et al., 2021) with *mapKurator*, a machine learning (ML) pipeline for automatic text detection and entity linking (Li et al., 2020). *MapKurator* suggests an initial set of text-bounding polygons, and users can accept, edit or delete these suggestions in *Recogito*. Further bespoke *Recogito* features enable one to capture a) how labels interact with each other and with visual elements (like colors and icons), and b) what semiotic functions labels perform (e.g. locative or complementary) (Schlichtmann, 2018). Structured data produced through this "deep annotation" are used to analyze geo-historical issues like industrialisation (Hosseini et al. 2021). Manually-annotated text data from maps may provide training data to improve and evaluate ML methods, but they also function as valuable datasets in their own right, particularly for smaller map corpora that can be annotated without recourse to ML.

Using our ML approach we predict 1) the *type* of content map text describes (i.e. roads, buildings, mountains) and, for unique features, 2) links in knowledge bases such as gazetteers or Wikidata. Through this process, we unlock the potential for users to find and interpret maps by the thousands based on search by semantic types. Using the links to specific instances of places, cultural institutions can feed this data back into their catalogs to document and study the geographical coverage of their collections. One could also explore differences between existing metadata and reported locations of map labels.

In a historical research case study, we use this method to analyze labels on large-scale British Ordnance Survey (OS) maps, investigating attitudes towards historical sites during the nineteenth-century (Fleet 2011), how maps communicate national historical narratives, and the fabrication of a common idea of "The Past" (Eggert 2009). Combining manual and automatic annotation provides rich information about the distribution of historical sites as types of places. Text data (including its location, spelling, fonts, and classifications) about historical sites enrich our understanding of the ways that early OS maps represented certain periods (Anglo-Saxon, Roman or Medieval). A diachronic analysis of the text labels offers initial answers about patterns in national-scale coverage of these cartographic features and prompts further questions about the historical, social, and cultural dynamics influencing the inclusion of antiquities on OS maps and their reception among the public.

Bibliography

Aucott, P., & Southall, H. (2019). Locating past places in Britain: creating and evaluating the GB1900 Gazetteer. *International Journal of Humanities and Arts Computing*, *13*(1-2), 69-94.

Eggert, P. *(2009). Securing the past: conservation in art, architecture and literature.*

Fleet, C. *(2011). Guest Editorial: Mapping and Antiquities in Scotland. Scottish Geographical Journal, 127(2), 85-86.*

Li, Z., Chiang, Y. Y., Tavakkol, S., Shbita, B., Uhl, J. H., Leyk, S., & Knoblock, C. A. (2020, August). An Automatic Approach for Generating Rich, Linked Geo-Metadata from Historical Map Images. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 3290-3298).

Hosseini, K., McDonough, K., van Strien, D., Vane, O., & Wilson, D. C. (2021). Maps of a nation? the digitized ordnance survey for new historical research. *Journal of Victorian Culture*, *26*(2), 284-299.

Schlichtmann, H. (2018). Background to the semiotic study of maps. *meta-carto-semiotics*, *11*(1), 1-12.

Vitale, V., Soto, P. D., Simon, R., Barker, E., Isaksen, L., & Kahn, R. (2021). Pelagios–Connecting Histories of Place. Part I: Methods and Tools. *International Journal of Humanities and Arts Computing*, *15*(1-2), 5-32.

The Postil Time Machine: The Lithuanian Lutheran Postils of the 16th Century

Chiarcos, Christian
christian.chiarcos@gmail.com
Goethe Universität Frankfurt, Germany

Gelumbeckaite, Jolanta
gelumbeckaite@em.uni-frankfurt.de
Goethe Universität Frankfurt, Germany

Drach, Mortimer
drach@em.uni-frankfurt.de
Goethe Universität Frankfurt, Germany

We introduce the Postil Time Machine, a novel research project dedicated to the study of knowledge transfer in 16th century Europe, with a particular focus on the situation in Lithuania. The project is funded by the German Research Foundation (DFG, 2021-2024) as a collaboration between Baltic studies and computational linguistics and pursues historical-philological as well as technological research questions.

In terms of philology, the project will provide the digital edition and linguistic annotation of the Old Lithuanian Lutheran postils and their Latin, German, and (in the case of the Bible passages) Greek and Hebrew sources, the detailed qualitative and quantitative analysis of the intertextual relations among the Lithuanian texts and of the relations between them and their respective foreign-language sources. The term 'postils' originally referred to Bible commentaries (Latin post illa verba textus "after these words from the scripture"), but later came to mean pericope sermons, and during the time covered by our project referred to an annual cycles of homilies. The study of postils and their transnational, theological and cross-lingual ties with political and theological movements (and the literature of said movements) at the time can provide a window into the formative period of Protestant faith. Furthermore, the postils covered here are among the oldest textual witnesses for the Baltic languages, and thus are of particular importance for the study of Lithuanian in the context of Indo-European studies and comparative linguistics (Gelumbeckaite, 2018).

In terms of technology, project goals include the development of a language technology stack for Old Lithuanian, the automated, cross-lingual detection of intertextual relations and citations, and the digital edition of the texts and their relations. While the first two aspects are innovative applications of existing technologies to a novel (and due to the sparsity and heterogeneity of available data, challenging) domain, the digital edition has somewhat larger implications for Digital Humanities as a whole: In the project we bring together four distinct types of data for different applications:

- manual annotation/IGT: Manual linguistic annotation is performed using the Field Linguist's Toolbox, standard software for the annotation of interlinear glossed text, which produces an idiosyncratic text format.
- automated annotation/CoNLL: The automated annotation is performed using state-of-the-art NLP software that produces CoNLL-TSV, a tabular data format. For training NLP tools, we convert Toolbox data to CoNLL.
- intertextual relations/Linked Data: Intertextual relations are manually annotated in Toolbox and automatically detected as part of the NLP workflow, but subsequently stored as a knowledge graph separate from the text and published as Linked Data. This facilitates interoperability with other initiatives working on intertextual relations and theological discourse.
- digital edition/TEI: The digital edition of text and annotations uses TEI and an XML stack.

While these solutions and the technologies behind are all well established, their combination has proven to be challenging on different levels, and this is where a major contribution of the project lies. These challenges include the limited interoperability between IGT and CoNLL formats, a lack of agreement about how to represent IGTs in TEI, and the conjunct usage of TEI and LOD technologies. Focusing on the latter aspect, we have been advocating the use of RDFa for this purpose (Tittel et al., 2018), as this is (at the moment) the only W3C-compliant way to implement a TEI-LOD bridge in inline XML (Chiarcos and Ionov, 2018). After more than a decade of discussion, this suggestion has eventually led to a novel TEI customization for TEI +RDFa (see https://github.com/TEIC/TEI/issues/1860). Our project represents the first application of this novel customization to a large-scale edition project and aims to demonstrate its scalability, robustness and, as a specific benefit, the application of off-the-shelf RDF technology to digital editions (as well as human-readable representations generated from it) created in this way.

Bibliography

Chiarcos, Ch., and Ionov, M. (2018), Linking the TEI: Approaches, Limitations, Use Cases. Paper presented at DH2019, Utrecht, The Netherlands.

Gelumbeckaitė, J. (2018). Predigtkultur in Litauen: Corpus der altlitauischen Postillen. In Reformatio Baltica (pp. 573-586). De Gruyter.

Tittel, S., Bermúdez-Sabel, H., & Chiarcos, C. (2018). Using RDFa to link text and dictionary data for Medieval French. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan.

Building infrastructure for annotating medieval, classical and pre-orthographic languages: the Pyrrha ecosystem

Clérice, Thibault
thibault.clerice@chartes.psl.eu
Centre Jean Mabillon, École nationale des Chartes, PSL, France

Jolivet, Vincent
vincent.jolivet@chartes.psl.eu
Centre Jean Mabillon, École nationale des Chartes, PSL, France

Pilla, Julien
julien.pilla@chartes.psl.eu
Centre Jean Mabillon, École nationale des Chartes, PSL, France

Introduction

For the past five years, we have been working on the development of infrastructure to build corpora and machine learning models for lemmatisation and morphosyntactic tagging. Ancient and medieval languages with rich morphology and high spelling variation represent a hanging fruit in the domain of these corpora. However, producing "gold" corpora is a tedious and costly task: even when the automatically produced annotations gain in quality thanks to ever more efficient models and vice versa, a significant amount of manual correction and validation work remains.

To reduce the cost and guarantee the interoperability of our corpora, we have built an ecosystem: (1) Pyrrha, a post-correction webapp for lemmatisation and morphosyntactic tags, (2) PyrrhaCI, a continuous integration tool for validating corpora, (3) Protogenie for merging and standardizing sometimes heterogeneous corpora, (4) Pie-Extended, a tagger taking into account the difference between real-world data training corpora and (5) Deucalion, a web service for annotation.

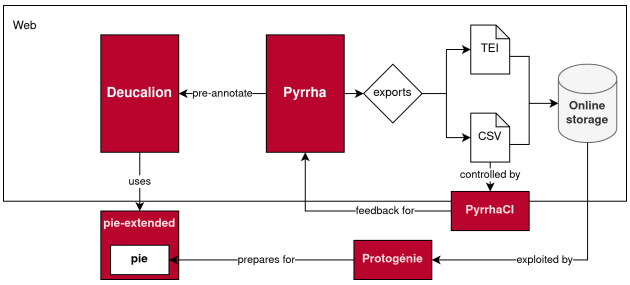


Figure 1:
Infrastructure developed at the École nationale des chartes

Producing data

Pyrrha (Clérice and Pilla, 2021) is designed to accelerate the correction of lemmatisation and morphosyntactic annotation. When we started our work, our team members were using spreadsheets, which have the ability to display all tags and context at the same time. The Pyrrha web application takes up this principle of a tabular interface but adds powerful validation functionalities thanks to checklists (lexicons of lemmas and morphosyntactic tags) guaranteeing the interoperability of the newly produced corpora, as well as batch correction functionalities, inspired by PoCoto (Vobl et al., 2014), a correction interface for OCR. Both of these functionalities are at the core of Pyrrha and have proven to be useful in speeding up the correction of out-of-domain corpora (*cf.* Figure 2). The application also allows collaboration, both for corpora and checklists, logging of corrections and export to multiple standards such as TEI and TSV.

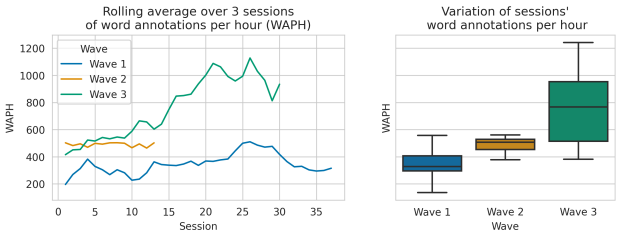


Figure 2:
Rolling average of the number of checked tokens per hour. Checking a token includes verifying its correctness and correcting it if necessary. Three waves of correction are visible, the first corpus was completely out-of-domain compared to the lemmatizer (Classical Latin), with the end of the corpus being very different from the rest of the corpus in terms of spelling (letters K and W appeared), themes and syntax. The second and third wave benefits of a new model, retrained on the data produced in wave 1: as a result, there is less corrections and a faster checking rate on waves 2 and 3. Wave 3 is composed of two blocs:

one is Thomas More's Utopia *(beginning of W3) and the other the Legenda Aurea which was nearly 100% correct, hence the effectiveness. Each wave / corpus was corrected and proofread on all categories that Pyrrha allows: Lemma, POS and morphological tags.*

Curating Corpora

While *Pyrrha* produces data that should be compliant with standard reference sets, mistake happens. *PyrrhaCI* (Clérice et al., 2021) is meant for testing the following attributes of datasets

1. respect of reference sets;
2. cross-categorical annotations (*e.g.* POS(dog)!=Verb);
3. n-gram tagging (*e.g.* ADJ should not be followed by VERcon).

Each test failure can be manually ignored for further tests, allowing for a more agile interpretation of grammar. PyrrhaCI is meant to be used as a continuous integration tool, through Github Actions or TravisCI, to validate datasets in open repositories and track the issues raised by editions.

Protogenie (Clérice, 2020b) is focused on preparing datasets for training. It is meant for the following:

1. keeping track of and using the same original train/dev/test splits while adding new data in order to have "uniform"' evaluation,
2. allowing for normalization of datasets that come from different projects in different formats,
3. adding transformation to the original dataset (while respecting (1)), such as removing the distinction between U and V in Latin, replacing labels, splitting multi-categorical tags, etc.

While (1) is easily taken care of, it is, in our experience, common to find datasets with different formatting choices or data-based variations such as punctuation, capitalization, morphological tags. Protogenie enables normalization of the "behavior" of different corpora, without having to work with pre-modified files, facilitating easy update of the latter and ensuring the stability of training and performances evaluation.

Producing new data: our lemmatization pipeline

Our models are trained with *Pie (Manjavacas et al., 2019)* [1], a lemmatizer with state-of-the-art results on pre-orthographic and morphologically rich languages and a relatively flexible and stable python API. Once trained, our models are served through *Pie-Extended*. Its first function is to bridge the gap between the real-world data and the curated training data by normalizing the first according to the latter [2]. It also provides features such as token [3] passthrough (*cf.* Figure 3).

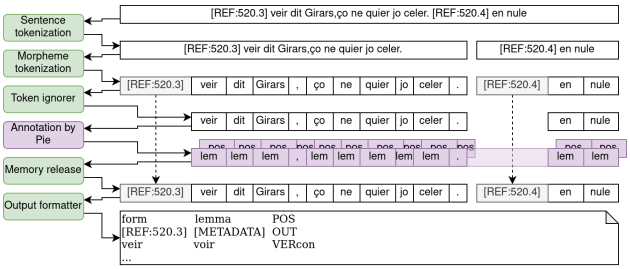


Figure 3:
Steps for token pass-through in Pie-Extended

Finally, as not everyone knows how to install and run a python program in a shell, we produced the Deucalion interface (Clérice, 2020a), meant both for documenting (with complete bibliography for each model and software) and tagging. It is a software layer allowing to use Pie-Extended on the web. This Deucalion interface can be used as a stand-alone web application or an API.

Conclusion

Over the last five years, this infrastructure has allowed us to build a 1+ million token corpus of Old French (Camps et al., 2021), a couple of datasets in both classical and late Latin (Clérice, 2021; Glaise and Clérice, 2021), one for pre-orthographic Early Modern French (Gabay et al., 2020), and others. There are improvements we would still like to make (*e.g.* the user-friendliness and capacities of PyrrhaCI) and we now have our eyes on *data valorization*, through the reuse of tools such as Blacklab (Does, de et al., 2017) [4]. Pyrrha has made lemmatization easier for our collaborators and made it a simpler task to produce data and share them across projects. This paper will be an opportunity to present a proven ecosystem, and also to assess its benefits, its costs and shortcomings.

Bibliography

Camps, J.-B., Clérice, T., Duval, F., Ing, L., Kanaoka, N. and Pinche, A. (2021). Corpus and Models for Lemmatisation and POS-tagging of Old French.

Clérice, T. (2020a). *Flask_pie, a Pie-Extended Wrapper for Flask*. https://github.com/hipster-philology/flask_pie.

Clérice, T. (2020b). *Protogenie, Post-Processing for NLP Dataset*. Zenodo doi:10.5281/zenodo.3883585. https://doi.org/10.5281/zenodo.3883585.

Clérice, T. (2021). Lemmatisation et analyse morpho-syntaxique des Priapées. https://github.com/lascivaroma/priapea-lemmatization.

Clérice, T., Blotière, É. and Schmied, M.-C. (2021). *PyrrhaCI*. https://github.com/hipster-philology/pyrrhaCI.

Clérice, T. and Pilla, J. (2021). *Pyrrha*. doi:10.5281/zenodo.2325427.

Does, J. de, Niestadt, J. and Depuydt, K. (2017). Creating research environments with blacklab. *CLARIN in the Low Countries*: 245–58.

Gabay, S., Clérice, T., Camps, J.-B., Tanguy, J.-B. and Gille-Levenson, M. (2020). Standardizing linguistic data: method and tools for annotating (pre-orthographic) French. *Proceedings of the 2nd International Conference on Digital Tools & Uses Congress*. pp. 1–7.

Glaise, A. and Clérice, T. (2021). Du IIème siècle à Thomas More, un corpus gold de latin lemmatisé et annoté en morpho-syntaxe. doi:10.5281/zenodo.1234. https://github.com/chartes/latin-non-classical-data.

Manjavacas, E., Kádár, Á. and Kestemont, M. (2019). Improving Lemmatization of Non-Standard Languages with Joint Learning. *ArXiv:1903.06939 [Cs]* http://arxiv.org/abs/1903.06939 (accessed 24 November 2019).

Vobl, T., Gotscharek, A., Reffle, U., Ringlstetter, C. and Schulz, K. U. (2014). PoCoTo - an Open Source System for Efficient Interactive Postcorrection of OCRed Historical Texts. *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*. (DATeCH '14). New York, NY, USA: ACM, pp. 57–61 doi:10.1145/2595188.2595197. http://doi.acm.org/10.1145/2595188.2595197 (accessed 26 November 2018).

Notes

1. We moved to a small fork of Pie, https://github.com/lascivaroma/PaPie, which includes tailored functionalities for our training sets.
2. *E.g.* in our Latin dataset shared with us by the LASLA, there was no punctuation. Unknown character can trigger weird behaviors in neural networks system, from our experience, creating issues for both the context of other lemma and its own lemmatization.
3. *E.g.* metadata token with text identifiers.
4. A demo for Latin is available at https://blacklab.alpheios.net/latin-texts/search thanks to Alpheios.

The Language of Reviewers: Sentiment, Ratings, and Style in Japanese-Language Amazon Video Reviews

Conroy, Melanie
mrconroy@memphis.edu
University of Memphis, United States of America

Nishi, Hironori
hnishi1@memphis.edu
University of Memphis, United States of America

Introduction

Online reviews have been analyzed for many traits that are linked to review quality: evidence of repetition, sarcasm, and of other markers of speech that may indicate the reviewer's insincerity, the use of bots, or a low-quality review (Wu, Van der Heijden, & Korfiatis, 2011; Li & Shimizu 2018; Lin & Kalwani, 2018). Within DH, Amazon reviews have been studied for what they tell us about the products reviewed, particularly books and their reception (Finn, 2011). We are interested in what the reviews can tell us about the process and style of reviewing, particularly in non-English languages. In this short talk, we offer some preliminary conclusions from our analysis of language, topics, and sentiments in video reviews in the Japanese-language portion of the Multilingual Amazon Reviews Corpus (Keung et al., 2020). Large-scale analyses of Amazon reviews have previously shown that there are fewer Japanese-language reviews on Amazon than English-language reviews (Lin & Kalwani, 2018); otherwise, reviews share many traits across languages (Keung et al., 2020).

Length of reviews

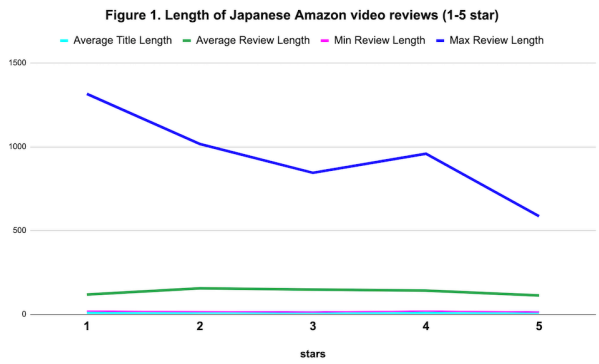


Figure 1 shows the average length of reviews, the max review length, the min review length, and the average number of characters in a review's title by star level in the 2,600 Japanese-language video reviews. Two- and three-star reviews are longer; many one-star reviews are quite short, or even vulgarly dismissive of the work being reviewed, yet some authors of one-star reviews feel the need to go into great detail about the lack of merit of the work reviewed (see max).

Topics of reviews

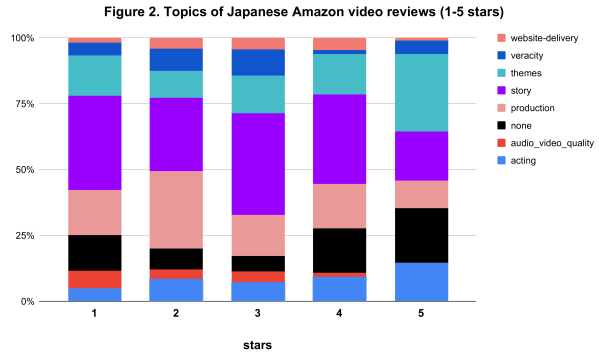
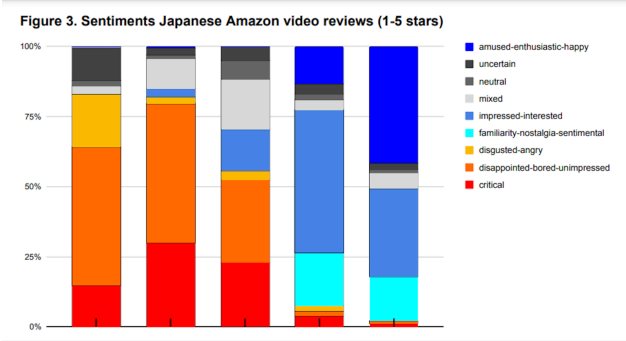


Figure 2 shows the distribution of main topics of Amazon video reviews. At all rating levels, there are diverse reasons for liking or disliking videos, as well as reviewers who failed to give any reason at all. The story was the most commonly referenced factor across all rating levels. Four-star and five-star reviews were more likely to reference acting or specific actors. One-star and two-star ratings more often referred to elements of production such as the adaptation, direction, editing, or cinematography, or problems with the audio or video quality. There were also many reviews referencing Amazon or the process of renting or buying the video in the mid-rated reviews. Five-star reviews often gave no reason for that rating or referenced themes.

Sentiments in reviews

We see both surprising and unsurprising patterns in the distribution of the predominant sentiment associated with each review, which was detected using key words and confirmed by a human reader, and which varied considerably (see figure 3). Three-star and lower reviews are more likely to be highly critical or express disappointment, boredom, or a failure to finish watching. One-star reviews are the most likely to express confusion or uncertainty about the quality or purpose of the video. Five-star reviews were mostly enthusiastic. Yet there is much evidence that the text of reviews does not align with the numerical rating given. A number of five-star reviews contain text that is mostly critical of the film. Critical reviews are more likely to be two-star or three-star reviews than one-star. Nostalgia or familiarity was the predominant sentiment in many of the reviews associated with the highest ratings.



Conclusions

While many reviewers' ratings and sentiment aligned, there were other reviews that gave no reasons for their rating and lots that expressed boredom or judgment without referencing qualities of the video. In other words, many reviewers were happy to express their approval or judgment of a video without feeling the need to justify their criticism. Some others gave very high numerical ratings but remained critical in the text of the review. Likewise, the praise for themes and actors in positive reviews and the frequent criticism of the crew and production in negative reviews were also notable. The presence of numerical ratings in this dataset allows us to compare the rhetoric of reviewers

and their ratings, as well as to locate reviews whose ratings do not align with the sentiment expressed in the review; such "misaligned" ratings can often tell us a lot about internet-based criticism of cultural works and which aspects of the work these critics target. Finally, we can compare these aligned and misaligned ratings across cultures to see whether critics in other cultures use the same rhetoric.

Bibliography

Finn, E. (2011). Reading, writing and reputation: literary networks in contemporary American fiction. *Digital Humanities 2011: Conference Abstracts. Stanford: Stanford University, pp.*47-49.

Keung, P., Lu, Y., Szarvas, G., & Smith. N. A. (2020). The multilingual Amazon reviews corpus. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 4563-4568.

Li, Z., & Shimizu, A. (2018). Impact of online customer reviews on sales outcomes: an empirical study based on prospect theory. *The Review of Socionetwork Strategies*, 12(2): 135-151.

Lin, H. C., & Kalwani, M. U. (2018). Culturally contingent electronic word-of-mouth signaling and screening: a comparative study of product reviews in the United States and Japan. *Journal of International Marketing*, 26(2): 80-102.

Wu, P. F., Van der Heijden, H., & Korfiatis, N. (2011). The influences of negativity and review quality on the helpfulness of online reviews. *International Conference on Information Systems*, 2011, 10 pages.

Scaling Up and Scaling Down: Expanding and Contracting in the Move to Linked Data

Crompton, Constance
constance.crompton@gmail.com
University of Ottawa, Canada

Schwartz, Michelle
michelle.schwartz@ryerson.ca
X University (formerly Ryerson University), Canada

Dangoisse, Pascale
pdang034@uottawa.ca
University of Ottawa, Canada

Lipski, Candice
candicelipski@outlook.com
University of Ottawa, Canada

Our project, Lesbian and Gay Liberation in Canada (LGLC, lglc.ca), has reached an exciting new phase: after creating 34,000 digital records documenting gay liberation in Canada from 1964 to 1981, we are expanding the project, both technologically and in scope. While popular histories of gay liberation in Canada tend to focus on the activism of gay, white, urban, anglophone men, the analysis of the data from the first six years of the LGLC project has shown that gay liberation in Canada was not solely driven by this demographic. This next phase of the LGLC project, started in 2020, combines archival research and linked data creation to represent the diversity and span of gay liberation organizing, intersectional personhood, and activist knowledge exchange, with a particular focus on women's and francophone activism (Crenshaw; Sheffeild; Gevisser; Bilge). The end date of the original project chronology, 1981, marks the start of the AIDS crisis, a milestone in Canadian gay men's organizing. This end date centres men's experience and did not, as our critics have pointed out, help the project meet its goals of documenting patterns in women's activism. We have expanded the project's end date to 1985 to capture significant organizing, and resulting social and legal change, by lesbian mothers and trans women.

This expansion of the project coincides with a new phase of our digital scholarship: the conversion of its content, drawn from archival sources, to linked data. While the project uses a graph data model akin to RDF, the strictures of existing ontologies create a flattening effect, as they do on most data derived from archives. Moreover, while archival research may reveal how people understood their historical situation and the identities available to or forged by them, the archival record is not neutral (Roberto; Drabinski). In our conversion to linked data we take our cue from Klein and D'Ignazio's call to preserve diverse (i.e. messy) data as a way to push back against statistical analysis' eugenicist roots, roots that have historically shaped cultural attitudes towards queer and other marginalised people (Klein and D'Ignazio; Brown; Maxwell; Tanguay). In this short paper we will share the results of our experiments in modeling our person data in CIDOC-CRM, chosen for maximum interoperability, and in a combination of Schema.org and the CWRC ontology, to most closely match our original data structure.

Bibliography

Bilge, Sirma. “Théorisations féministes de l’intersectionnalité.” *Diogène*, vol. 225, no. 1, Presses Universitaires de France, 2009, pp. 70–88.

Brown, Susan. “Categorically Provisional.” *PMLA*, vol. 135, no. 1, Modern Language Association, 2020, pp. 165–74. doi:10.1632/pmla.2020.135.1.165.

Crenshaw, Kimberle. “Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics.” *University of Chicago Legal Forum*, vol. 1989, 1989, pp. 139–68.

D’Ignazio, Catherine, and Lauren F. Klein. *Data Feminism*. MIT P, 2020.

Drabinski, Emily. “Queering the Catalog: Queer Theory and the Politics of Correction.” *The Library Quarterly: Information, Community, Policy*, vol. 83, no. 2, 2013, pp. 94–111. doi:10.1086/669547.

Gevisser, Mark. *The Pink Line: Journeys Across the World’s Queer Frontiers*. Farrar, Straus and Giroux, 2020.

Maxwell, John. “Text Processing Techniques and Traditions (or: Why the History of Computing Matters to DH).” *Doing More Digital Humanities: Open Approaches to Creation, Growth, and Development*, edited by Constance Crompton, Raymond Siemens, and Richard Lane., Routledge, 2019.

Roberto, K. R. *Radical Cataloging: Essays at the Front*. McFarland, 2008.

Sheffield, Rebecka T. “More than Acid-Free Folders: Extending the Concept of Preservation to Include the Stewardship of Unexplored Histories.” *Library Trends*, vol. 64, no. 3, Johns Hopkins UP, 2016, pp. 572–84. doi:10.1353/lib.2016.0001.

Tanguay, Christian. “Les archives LGBTI: le droit à un lieu pour sa culture.” *Les cahiers de la LCD*, vol.1, no. HS1, 2018, pp. 145–54.

Ware, Syrus Marcus. *“No One Like Me Seemed to Have Ever Existed”: A Trans of Colour Critique of Trans Scholarship and Policy Development in Post-Secondary Schools*. University of Toronto, 2011.

Una mirada digital a cuatro obras dramáticas de Federico García Lorca. Aportes de las técnicas de visualización a la interpretación de texto dramático

Dabrowska, Monika
monika.dabrowska@unir.net
Universidad Internacional de la Rioja (UNIR)

Santa Maria, Teresa
teresa.santamaria@unir.net
Universidad Internacional de la Rioja (UNIR)

Introducción

La propuesta parte del convencimiento del gran potencial de la visualización de la información para el estudio de los textos literarios, apoyándose en la teoría de las redes sociales y técnicas de la visualización de los datos. Si el texto literario es una red de fenómenos discursivos, lo es por excelencia una obra dramática, una red de parlamentos e interacciones entre los personajes. Aplicando los instrumentos de análisis y representación gráfica, se pueden ilustrar las relaciones discursivas entre los personajes. El tratamiento de los datos permite visualizar y “medir” estas relaciones, así como evidenciar fenómenos difícilmente perceptibles en la lectura tradicional, creando nuevos conocimientos e interpretaciones.

Objetivo

El propósito de esta presentación es mostrar ejemplos de visualización de texto y sus posibilidades interpretativas, basándose en los cuatro dramas rurales del autor granadino: *Bodas de sangre* (1933), *Yerma* (1934), *Doña Rosita la soltera* (1935) *La casa de Bernarda Alba* (1936). Se exploran las potencialidades de Gephi y RAWGraphs para sintetizar, sistematizar y representar los datos mediante diferentes grafos y otras representaciones infográficas (dendrogramas, diagramas aluviales, etc.). La exploración muestra los nuevos conocimientos que se pueden extraer a partir de transformación de los datos en imágenes, referentes a la estructura, composición y contenidos de las obras teatrales.

Método

En el trabajo se aplican métodos de análisis cuantitativo y cualitativo. Los textos digitalizados en formato TEI, GEFX y CVS provienen de Drama Corpora Proyect (DraCor, disponible en:). Para representar la red de conexiones entre los personajes se crearán los grafos, que representan la red de protagonistas y el peso cuantitativo de las intervenciones discursivas de cada una. Para su procesamiento y generación de representaciones gráficas se utiliza el software libre Gephi 0.9.2-beta (Gephi.org 2008-2017) y RAWGraphs (https://rawgraphs.io/). Se comparan los grafos generados por ambas herramientas con los creados por la API de DraCor llamada Shiny DraCor (). Para visualizar las relaciones entre las palabras y unidades léxicas más significativas se recurre a los grafos léxicos. Para mapear las estadísticas textuales se rastrean otras 29 herramientas de representación gráfica incluidas en la interfaz de RAWGraphs.

Avance de conclusiones

El interés principal de la propuesta radica en la exploración de las visualizaciones y detección de las singularidades de los elementos léxicos en el tejido del texto lorquiano. Los resultados de análisis informático se confrontan con los enfoques críticos tradicionales. Se evalúa críticamente la eficacia de las técnicas de visualización para la interpretación de la obra literaria, su valor epistemológico, las posibilidades y limitaciones de las herramientas utilizadas (algunas de ellas automatizadas) y los retos que planean a los estudios literarios. Sin duda, es una vía que puede revelar tendencias y patrones inesperados y aportar claves interpretativas de interés.

Bibliografía

Barros García, B. B. (2020). El texto literario hecho datos: F. M. Dostoievski en el marco de las Humanidades digitales y los enfoques cuantitativos. *452ºF. Revista De Teoría De La Literatura Y Literatura Comparada*, (23), 53–77.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proceedings of the International AAAI Conference on Web and Social Media*, *3*(1), 361-362.

Burley, D., Ashburn, V. (2010). Information Visualization As A Knowledge Integration Tool. *Journal of Knowledge Management Practice*, 11(4), 18.

Cherven, K. (2015). *Mastering Gephi Network Visualization*. Packt Publishing Ltd

Drucker, J. (2011). Humanities Approaches to Graphical Display. *Digital Humanities Quarterly*, 5 (1)

Martínez Carro, Elena. (2018) Los personajes femeninos lorquianos desde una interpretación digital, en: Álvarez Ramos, Eva y Blasco Pascual, Javier (eds.). *Humanidades Digitales. Retos, Recursos y Nuevas Propuestas,* Agilice Digital, Valladolid.

Santa María, M.T., Calvo Tello, J., Jiménez, C.M. (2021). ¿Existe correlación entre importancia y centralidad? Evaluación de personajes con redes sociales en obras teatrales de la Edad de Plata, Digital Scholarship in the Humanities, Volume 36 (1), 81–88. https://doi.org/10.1093/llc/fqaa015

Word frequencies in authorship attribution: A simple tweak to improve performance

Eder, Maciej

maciej.eder@ijp.pan.pl
Institute of Polish Language (Polish Academy of Sciences)

Abstract

In this paper, I introduce a very simple method of computing relative word frequencies for authorship attribution and similar stylometric applications. The proposed method outperforms classical most-frequent-word approaches by a few percentage points.

Introduction

In a vast majority of stylometric studies, relative frequencies of the most frequent words (MFWs) are used as the language features to betray the authorial "fingerprint". A vector of such relative word frequencies is then passed to one of the multidimensional machine-learning classification techniques, ranging from simple distance-based lazy learners, such as Delta (Burrows, 2002; Evert et al., 2017), to sophisticated neural network setups (Gómez-Adorno et al., 2018).

Even if alternative types of features have been introduced (Peng et al., 2003; Hirst and Feiguina, 2007; Lučić and Blake, 2013) and tested in controlled experiments (Eder, 2011), the standard approach relying of word frequencies continues to be predominant in the field (Grieve, 2007; Stamatatos, 2009). In this paper, I propose to count the relative frequencies in a slightly different way, in order to better capture the authorial choice of words.

Words that (might) matter

The notion of relative word frequencies is fairly simple. We count all the tokens belonging to particular types (e.g. all the attestations of the word "the", followed by the attestations of "in", "for", "of" etc.), and for each word, we divide the number of types by the total number of words in a document. Consequently, each word frequency is equal to its percentage within the document (e.g. "the" = 0.0382), and all the frequencies sum up to 1. The reason of converting occurrences to relative frequencies is obvious: by doing so, one is able to reliably compare texts that differ in length.

For the sake of this paper, it is important to note that such frequencies are relative to *all the other words* in a document in question. Convenient as they are, these values are at the same time very small and – importantly – can be affected by other word frequencies. Now, what if we disregard thousands of other words in a text, and compute

the frequencies in relation to a small number of words that are *relevant*? An obvious example is the mutual relation between the words "on" and "upon" in one document (Mosteller and Wallace, 1964); essentially, more attestations of "upon" comes at the cost of "on" being less frequent, and vice versa. While the classical relative frequency of the word "on" in Emily Bronte's *Wuthering Heights* is 0.00687, the proportion of "on" relative exclusively to "upon" is 0.9762. It is assumed in this paper that the latter frequency can betray the authorial signal to a greater extent than the classical approach, because the myriads of other words are not blurring the final value.

Given the above assumption, it would be tempting to identify one synonym for each of the words, and to compute the relative proportions in each of the synonym pairs (Borski and Kokowski, 2021). Linguistically speaking, however, such an approach would hardly be feasible. Firstly, only a fraction of words have their synonyms. Secondly, some semantic fields are rather rich and cannot be reduced to a mere pair of synonyms. Thirdly, in the case of the most frequent words (articles, particles, prepositions) seeking their synonyms doesn't make much sense, yet still, *relevant counterparts* for these frequent words obviously exist. Rather than identifying rigid lexical synonyms, then, I used a word embedding model (GloVe, 100 dimensions) to extract *n* semantic nearest neighbors for each of the words in question. Consequently, the neighbors for the word "person" were: "woman", "gentleman", "man", "one", "sort", "whom", "thing", "young", etc., whereas the neighbors for the word "the" were as follows: "of", "this", "in", "there", "on", "one", "which", "its", "was", "a", "and", etc. For each target word, a relative frequency was calculated as the number of occurrences divided by the sum of occurrences of its *n* semantic neighbors (*n* being the size of semantic space to be tested).

Results

In order to corroborate the above intuitions, a controlled experiment was designed. A benchmark corpus of 100 English novels (33 authorial classes) was used, together with the package 'stylo' to perform the tests (Eder et al., 2016). Different classifiers, MFW vectors and, most importantly, different sizes of the semantic space were tested systematically, in a supervised setup with stratified cross-validation. On theoretical grounds, the size of the semantic space $n = 37{,}000$ (roughly the total number word types in the benchmark corpus) would be equivalent to classical relative frequencies, whereas the space of the size $n = 1$ means that the frequencies are relative to exactly one other word (e.g. the frequency of the word "the" would be the number of occurrences of "the" divided by the total number of "the" and "of").

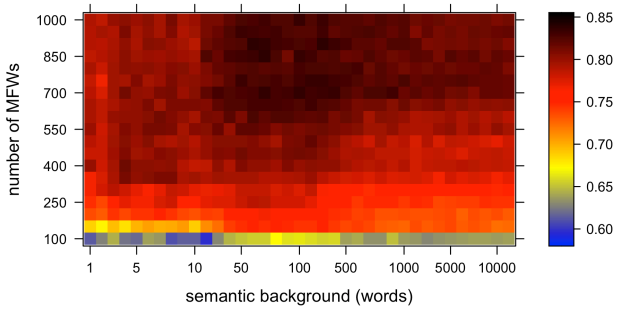


Figure 1:
The performance (F1 scores) for a benchmark corpus of English novels and Cosine Delta as a classifier. The results depend on the MFW vector (y axis) and the size of the semantic space (x axis). Classifier used: Classic Delta.

The obtained results clearly suggest that the new method outperforms the classical relative frequencies solution. In agreement with several previous studies, longer MFW vectors worked better than, say, 100 MFWs. Less intuitive was the fact that the best performing word frequencies were the ones relative to ca. 50 neighboring words (Fig. 1). A recipe for a robust authorship attribution setup seems to be as follows: take 1,000 MFWs, and compute their frequencies using, for each word, the occurrences of their 50 semantic neighbors.

Since authorship attribution results are proven to be unevenly distributed across different MFW vectors, Fig. 2 shows the performance of the model as the gain (in percentage points) over the standard solution. While the overall best performance is obtained for 1,000 MFWs and the space of 50 words, the biggest gain over the baseline (more than 5 percentage points) is produced by the vector of 100 MFWs, each of them computed as a frequency relative to its 80 neighboring words. Interestingly, the new method proves to be *worse* than the baseline for long MFW vectors and tight semantic spaces of 1–10 neighboring words.

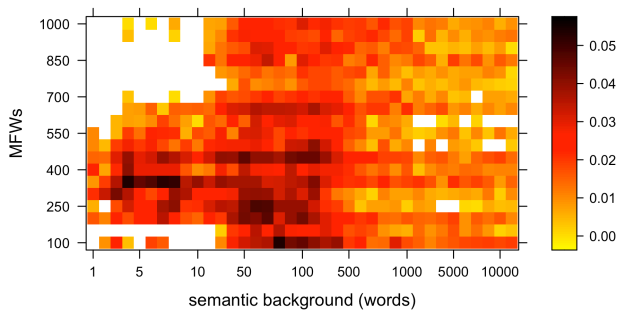


Figure 2:
The gain (in percentage points) over the classical relative frequencies, for different MFW vectors (y axis) and the size of the semantic space (x axis). Classifier used: Classic Delta.

Conclusion

The paper presented a simple method to improve the performance in different stylometric setups. The method is conceptually straightforward and does not require any NLP tooling. The only external piece of information that is required is a list of semantically related words for each of the most frequent words in the corpus.

Acknowledgements

This research is part of the project 2017/26/E/HS2/01019, supported by Poland's National Science Centre.

Bibliography

Borski, G. and Kokowski, M. (2021). Copernicus, his Latin style and comments to Commentariolus. *Studia Historiae Scientiarum*, **20**: 339–438.

Burrows, J. (2002). "Delta": A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**(3): 267–87.

Eder, M. (2011). Style-markers in authorship attribution: A cross-language study of the authorial fingerprint. *Studies in Polish Linguistics*, **6**: 99–114.

Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R: A package for computational text analysis. *R Journal*, **8**(1): 107–21.

Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C. and Vitt, T. (2017). Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities*, **32**(suppl. 2): 4–16 doi: 10.1093/llc/fqx023.

Gómez-Adorno, H., Posadas-Durán, J.-P., Sidorov, G. and Pinto, D. (2018). Document embeddings learned on various types of n-grams for cross-topic authorship attribution. *Computing*, **100**(7): 741–56 doi: 10.1007/s00607-018-0587-8.

Grieve, J. W. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary And Linguistic Computing*, **22**(3): 251–70 doi: 10.1093/llc/fqm020.

Hirst, G. and Feiguina, O. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, **22**(4): 405–17.

Lučić, A. and Blake, C. L. (2013). A syntactic characterization of authorship style surrounding proper names. *Digital Scholarship in the Humanities*, **30**(1): 53 doi: 10.1093/llc/fqt033.

Mosteller, F. and Wallace, D. (1964). *Inference and Disputed Authorship: The Federalist*. Stanford: CSLI Publications.

Peng, F., Schuurmans, D., Keselj, V. and Wang, S. (2003). Language independent authorship attribution using character level language models. *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*. pp. 267–74.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, **60**(3): 538–56.

US-CHINA ZIGZAG: Semantic shifts and sentiment analysis of the US Congress speeches on US policy towards China and Taiwan

Elkobi, Jonathan
Jonathan.Elkobi@mail.huji.ac.il
The Hebrew University of Jerusalem, Israel

The relations between The United States, The Republic of China (ROC), and The People's Republic of China (PRC), is a classic case of "there and back again" shifts in the international policy of the hegemon in the contemporary era. In this paper, I examine how the US addressed different Chinese entities at different times. Looking at a specially-curated corpus of U.S. congressional speeches regarding ROC and PRC and using a contextual language model, BERT (Devlin et al. 2018), to conduct sentiment analysis.

And a word2vec to examine semantic shifts (Hamilton et al. 2016). I will trace the changes in US policy toward the "Two Chinas" and examine when those shifts started to happen.

Short background: After the Chinese Communist Party won the civil war, The Western world saw ROC as the rightful heir of China; thus, ROC was a member of the UN while, and PRC was left behind under the one-china policy (Spence 1990). Few decades later, PRC and the US have moved to close ties under Nixon, with Kissinger leading the path for a warm relationship between the CCP and the US leading to ROC getting ousted from its place in the UN being replaced with PRC by 1971. In 1979 the US officially switched its recognition from ROC to PRC (Westad 2012). In the 21st century, the US discourse shift back to supporting ROC, and in 2007 the congress declared they do not recognize PRC's sovereignty over Taiwan (Kan 2007). Meanwhile, PRC had been growing expeditiously its economic power, they joined the WTO in 2001, and became a main player in the world economy. This brought more fear into the US congress. Peaking when the Trump administration started the trade war with PRC and continues until this day (Kwan 2020).

In my research, I will measure the semantic shifts in US discourse and the changes in sentiment analysis using a BERT model. The US relations with PRC and ROC are complicated, with double-crossing from the US side regarding the one-China policy. Thus, semantic shifts research of crucial points during the modern history of the relations can help identify the concrete shifting process from ROC to PRC and back again. The proposed research is unique due to its methodological use of combining semantic shifts with sentiment analysis to expose the trends in the discourse itself (Azarbonyad et al. 2017). By using these computational tools, we can find the meaning between the lines of the US Congress members in the crucial years that led to the shifts in the US policy regarding the one-china policy.

Data:

For this research, I will use the corpus of U.S. Congress speeches from 1949 to 2011 (Gentzkow et al. 2018).

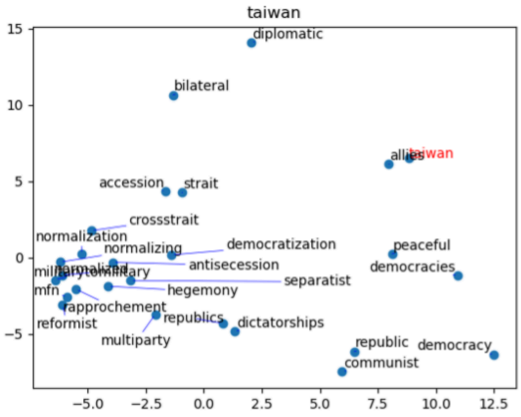
Methodology:

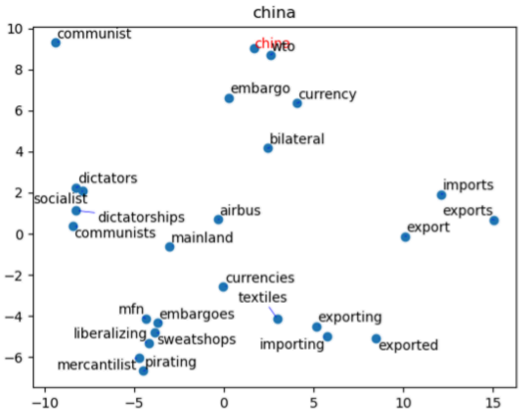
1. Semantic Shifts: I compute the word similarities through time to detect the semantic shifts and the semantic surrounding of PRC and ROC in the discourse of the US congress in those critical years, trying to point out when exactly those shifts occurred to connect them with the relevant events. Applying computational methods to measure semantic shifts affects the way scholars can interpret and validate theories about changes in the way the wind blows in every text-based discipline. Semantic shifts analysis is the state-of-the-art method for understanding the transformation of terms meaning throughout time by the word context, thus, we can follow those shifts in meaning to find how does and how the discourse has shifted regarding the terms "China", "Taiwan" and their equivalents (Kutuzov, et al. 2018; Zhang et al. 2016).
2. Sentiment Analysis: A second step will measure the semantic sentiment toward these two entities over time to correlate with the shifts in the meaning of words. This is a powerful tool to assess and measure the opinions inside textual forms for comprehending the writer's attitude towards the subject of the text (Feldman 2013). I will use aspect-based sentiment analysis on a BERT model (Xu et al. 2019) to quantify the attitude of the congress toward PRC and ROC over time, pointing to the shifts of the policy stance of the US towards those entities, using the terms terms "China", "Taiwan" and their equivalents.

Using those NLP tools will give us a measurable context to understand when the US discourse shifted to enable those switches in US policy. This research can lead us to comprehend deeply the US changing policies in the Pan-Chinese era about those two Chinese entities, when they started, by who, and in what magnitude.

Preliminary results based on semantic change using word2vec show shifts in the conception of the PRC in the US congress discourse. First, since 1949 until the 1959, the PRC is seen as a communist and hostile. Then it changed to a more accepting approach, followed by a change to a warmer approach of the PRC in 1971. Then a decline in friendliness after the 1989 Tiananmen events. And finally, it becomes a major economic player in the conception of the US. Meanwhile, the discourse regarding Taiwan would moves in the opposite direction, starting as an ally in the 1950, but receiving a colder shoulder after 1971, and lastly returning to be a democratic ally after Tiananmen.

Example of the top 25 nonentity terms in plot of the word2vec models for the terms 'China' and 'Taiwan' based on the Congress speeches between 2001- 2011:





Bibliography

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. "Diachronic word embeddings reveal statistical laws of semantic change." *arXiv preprint arXiv:1605.09096* (2016).

Spence, Jonathan D. *The search for modern China*. WW Norton & Company, 1990.

Westad, Odd Arne. *Restless empire: China and the world since 1750*. Hachette UK, 2012.

Kan, Shirley A. "China/Taiwan: Evolution of the" One China" Policy-Key Statements from Washington, Beijing, and Taipei." LIBRARY OF CONGRESS WASHINGTON DC CONGRESSIONAL RESEARCH SERVICE, 2007.

Kwan, Chi Hung. "The China–US trade war: Deep-rooted causes, shifting focus and uncertain prospects." *Asian Economic Policy Review* 15, no. 1 (2020): 55-72.

Azarbonyad, Hosein, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. "Words are malleable: Computing semantic shifts in political and media discourse." In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 1509-1518. 2017.

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts. Palo Alto, CA: Stanford Libraries [distributor], 2018-01-16. https://data.stanford.edu/congress_text

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. "Diachronic word embeddings and semantic shifts: a survey." *arXiv preprint arXiv:1806.03537* (2018).

Zhang, Yating, Adam Jatowt, Sourav S. Bhowmick, and Katsumi Tanaka. "The past is not a foreign country: Detecting semantically similar terms across time." *IEEE Transactions on Knowledge and Data Engineering* 28, no. 10 (2016): 2793-2807.

Feldman, Ronen. "Techniques and applications for sentiment analysis." *Communications of the ACM* 56, no. 4 (2013): 82-89.

Xu, Hu, Bing Liu, Lei Shu, and Philip S. Yu. "BERT post-training for review reading comprehension and aspect-based sentiment analysis." *arXiv preprint arXiv:1904.02232* (2019).

What can stylometry and LIWC infer from Octavia E. Butler's use of 6+ letter words?

Essam, Bacem
literaryartrans@gmail.com
Peerwith, Netherlands, The

The left temporal and temporo-parietal structures govern the memorization of words of the mental lexicon, while left frontal structures control the processing of the mental grammar and the computation of standard morphological forms (Ullman et al., 2005). This study compiles a corpus of Butler's twelve novel to be categorized morphologically and cognitively using Linguistic Inquiry Word Count (LIWC) program; phonetically using phonotactic probability; and R package for clustering phonemes, morphemes and lexemes. The phonotactic probability of all words was measured using KU's phonotactic probability calculator (Vitevitch, & Luce, 2004). Corpus tools were used to generate 5-word-window concordance for all sentences

containing a 6+ letter word. The literary section of Brown corpus that includes fiction written by non-dyslexic writers was used as a reference corpus. Brown corpus is freely searchable through Sketch Engine and is accessible for downloading. We also compiled another contemporary reference for contemporary non-dyslexic American writers as a second reference corpus. De Luca et al. (2008) suggested that word length selectively influenced word recognition in impaired versus skilled readers, regardless of the moderating action of sublexical, lexical, and semantic factors. In this study, the linguistic analysis of the concorded lines retrieved from Butler's novels, aims to test whether Butler's reading deficits affected her selection of lengthy words. The categorization of 6-letter and 9+letter words demonstrated that most lengthy words were either nominalized or jargon terms. The use of lengthy verbs was, however, inconspicuous. The phonotactic probability of the most frequently used words was complex in nouns and modifiers but not in verbs. The study suggests that Butler's choice of lengthy words is a coping mechanism she used to regulate her image as a dyslexic writer. Although De Luca et al. (2008) conducted a reading-task study, the reading and comprehension are neuro-linguistically dependent on the morphological complexity of the written text. Crisp et al (2011) concluded that the severely impaired phonology and the representation of semantic-phonological impairment demonstrate that semantic representations are central to reading in the face of phonological impairment. Implications inferred from this experiment are ushered towards effective computation of word imageability, frequency of use, selection of word senses, word familiarity, word length, semantic diversity and phonological neighborhood density.

Bibliography

Crisp, J., Howard, D., & Lambon Ralph, M. A. (2011). More evidence for a continuum between phonological and deep dyslexia: Novel data from three measures of direct orthography-to-phonology translation. Aphasiology, 25(5), 615-641.

De Luca, M., Barca, L., Burani, C., & Zoccolotti, P. (2008). The effect of word length and other sublexical, lexical, and semantic variables on developmental reading deficits. Cognitive and Behavioral Neurology, 21(4), 227-235.

Ullman, M. T., Pancheva, R., Love, T., Yee, E., Swinney, D., & Hickok, G. (2005). Neural correlates of lexicon and grammar: Evidence from the production, reading, and judgment of inflection in aphasia. Brain and Language, 93(2), 185-238.

Vitevitch, M.S. & Luce, P.A. (2004) A web-based interface to calculate phonotactic probability for words and nonwords in English. Behavior Research Methods, Instruments, and Computers, 36, 481-487.

States of the text: fixation, legitimation and multimodal publication

Fauchié, Antoine
antoine.fauchie@umontreal.ca
Chaire de recherche du Canada sur les écritures numériques, Canada

Publications (books, articles, etc.) are artifacts that carry texts, and their production requires complex editing processes, calling upon different types of skills and tools. Since the beginnings of computing and following a long technical evolution, digital technology has introduced mechanisms that modify both the editorial work and the workings necessary for the production of different forms and formats (Mourat, 2020). These technological *innovations*, which are numerous and sometimes reproduce pre-digital gestures identically (Ludovico and Cramer, 2012), are created and adopted in different stages or phases (Epron and Vitali-Rosati, 2018): theoretical principles are conceptualized, imagined and then implemented in prototypes, and finally industrialized for restricted or broad uses.

We are interested here in two technical principles derived from digital publishing technologies (Blanc and Haute, 2018), namely multimodal publishing and single source publishing. On the one hand, it is about being able to generate several different formats within the same editorial project — such as a printed book on paper, a digital book in EPUB format and an XML file —, rearranging the contents according to the form of the artifact, whether in terms of structure or graphic rendering (Haute, 2019). On the other hand, it is a question of producing these versions with a single source, limiting the usually repeated interventions on several source files, and making the editorial follow-up more fluid (Hyde, 2021). These two principles — both theoretical, implemented and proven — aim to simplify work on a text in the context of an editing process: to facilitate multiple interventions by the different people who act on a text, and thus to allow interventions at different levels and moments in the text production chain. These *states* of the text, whether reading, correcting, rereading, formatting, structuring, composing, transforming, publishing, can thus be considered and composed in various ways, without the risk of losing information or creating unnecessary conflicts between the participants.

If technological evolutions allow us to imagine new ways of working in publishing, what about their real implementation? The principles mentioned are implemented through methods, software or computer programs, often put forward to the detriment of questioning the organizations of humans who use these technologies (Gelgon, 2018). Can the legitimization of content (Vitali-Rosati, 2012) be integrated into a *single source publishing* chain? If multiple people can be involved in an editorial project simultaneously, text legitimation poses a significant challenge. At what point is a text fixed? If the contents can be modified at any time, and even technically after publication, the fixation of the text is questioned, as well as its citability (the text is moving) and its durability (several versions and states of the text coexist without systematic archiving). Finally, if the technical processes are numerous to realize this type of publication chain (Maxwell, 2019), perhaps they are not all desirable. We need to take a critical look at technical feasibility. Isn't a multimodal publishing chain based on a single source a utopia in the publishing ecosystem?

Our presentation focuses on the states of the text in relation to human organizations and technical solutions. We wish to analyze the mechanisms of fixation, legitimization and publication in the perspective of editing a text in several versions from a single source. Through several examples of publication chains, we will question the technical solutions at work, and we will draw up a panorama of the theoretical questions that are part of the digital humanities approach.

Bibliography

Blanc, J. and Haute, L. (2018). Technologies de l'édition numérique. *Sciences Du Design*, **8**(2): 11–17 https://www.cairn.info/revue-sciences-du-design-2018-2-page-11.htm (accessed 23 March 2019).

Epron, B. and Vitali-Rosati, M. (2018). *L'édition à L'ère Numérique*. (Repères). Paris, France: La Découverte https://www.cairn.info/l-edition-a-l-ere-numerique--9782707199355.htm.

Gelgon, A. (2018). Un dialogue à réaliser : Design et technique. In, *.Txt 3*. Paris, France: Éditions B42, p. 158.

Haute, L. (2019). Livres mécaniques et chimères numériques. *Back Office*(3) http://www.revue-backoffice.com/numeros/03-ecrire-lecran/06_haute.

Hyde, A. (2021). Single Source Publishing *Coko* https://coko.foundation/single-source-publishing/ (accessed 17 November 2021).

Ludovico, A. and Cramer, F. (2012). *Post-Digital Print: The Mutation of Publishing Since 1894*. Eindhoven, Pays-Bas: Onomatopee.

Maxwell, J. W. (2019). *Mind the Gap: A Landscape Analysis of Open Source Publishing Tools and Platforms*. Cambridge, Massachusetts, États-Unis d'Amérique: The MIT Press https://mindthegap.pubpub.org/ (accessed 30 December 2019).

Mourat, R. de (2020). Le vacillement des formats. Matérialité, écriture et enquête : Le design des publications en Sciences Humaines et Sociales Université Rennes 2 Thèse http://www.these.robindemourat.com/ (accessed 22 October 2020).

Vitali-Rosati, M. (2012). Auteur ou acteur du web ? *Implications Philosophiques* https://papyrus.bib.umontreal.ca/xmlui/handle/1866/12980.

An experimental attempt to use Transfer Learning for Named Entity Recognition in letters from the 19th and 20th century

Flüh, Marie
marie.flueh@uni-hamburg.de
Universität Hamburg, Germany

Lemke, Marc
marc.lemke@uni-rostock.de
Universität Rostock, Germany

In this contribution, we investigate to what extent data from one digital scholarly edition project (*Dehmel digital* [1]) can be used to fine-tune a large-scale language model, which was pre-trained for the purposes of another project (*The Complete Works of Uwe Johnson* [2]). We discuss the opportunities and practical limitations of such an attempt dealing with differences in the language properties of the material that was used to pre-train the transferred model and the material that is applied to fine-tune this model for the specific NER task. As part of this, we investigate to what extent the quantity and quality of training data affects the performance of NER models. This places our contribution in the line of approaches dealing with computational analysis of textual material from different periods of time (Labusch et al., 2019; Schmidt et al., 2021; Ehrmann et al. 2022).

Technical prerequisites

We use the software *NEISS TEI Entity Enricher* (NTEE). [3] Following a Transfer Learning (Kamath et al., 2019) approach the tool provides access to large-scale language models in a Bidirectional Encoder Representations from Transformers (BERT) architecture (Devlin et al., 2019; Underwood, 2019) for different languages, which can be

fine-tuned to be applicable for NER tasks. The pre-trained model selected for our investigation is a representation of modern German of the 20th and 21st centuries, trained on 8 GB of text data from German Wikipedia and a web crawl of various German newspaper portals (Zöllner et al., 2021: 2–3).

Datasets

The investigation is based on two differently sized datasets, which were taken from the *Dehmel digital* project and consist of German-language letters from the early 20th century (see table 1). 10 per cent of the sets were taken for validation purposes respectively.

Token category	big	small
Person	9397	3683
Place	3056	1092
Organisation	635	269
Work	654	294
unlabeled	190970	52271
total	204712	57609

Table 1: Composition of the big and small set of training data used as ground truth for NER model training

Assumptions

Historical-linguistics insights show that grammatical properties of language change slowly over long periods (Nübling et al., 2017). We assume that the difference in time between the 21st and the early 20th century is not associated with an essential difference in the German language system, that would prevent useful entity predictions.

We suppose that the difference at hand in text types (newspaper and encyclopedia articles vs. letters) can be neglected for our purpose based on the experiences made in the *Complete Works of Uwe Johnson* project: The same language model has already been successfully used for NER tasks on a corpus of letters.

Investigation

The investigation focuses on the question of how the amount of training data affects the performance of NER models for the detection of places, persons, organisations and works in historical letters.

The evaluation of the performance happens in two steps: We evaluate the E-F_1 of each training epoch determined on the validation set by NTEE itself during the training processes and we use the created models to predict entities on four specific sample texts, which are not part of the validation or the training set. Subsequently, we calculate Precision, Recall and E-F_1 for all of them. To determine the strengths and weaknesses in detail this quantitative approach is accompanied by a qualitative analysis of the predictions.

Selected results

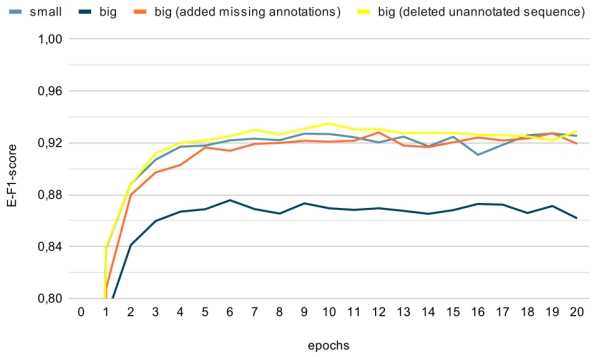


Figure 1:
E- *F 1-scores of four models comparing the performances depended on the ground truth data amount and annotation consistency*

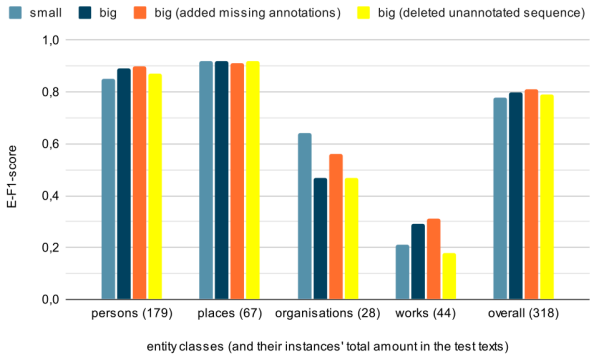


Figure 2:
E- *F 1-scores of four models calculated in the sample text prediction analysis, entity-wise and overall*

The big model performs worse on the validation set of its respective ground truth than the small one (fig. 1). This finding contradicts the general assumption, that more ground truth leads to better prediction results. But if we take a closer look at the data, we encounter a problem inside the big ground truth dataset: About 5.3% of the text was not annotated. In this case, fixing the big ground truth by adding the missing annotations or by deleting the unannotated

sequence leads to equivalent performance values to just using the small one.

The data collected leads to qualitative insights regarding the training process and the difficulties that come along with it. From the incorrectly predicted annotations, three problem categories can be derived:

1. a lack of ground truth data, including a lack of samples for representatives of the categories 'work' and 'organisations' within the training data,
2. inconsistent annotations,
3. ambiguous entities, which can belong to several categories.

In the talk, we would like to go into more detail on the quantitative and qualitative analyses to present conclusions with which we shed light on the computer-assisted, transfer-learning-based analysis of historical letters in digital scholarly editions.

Bibliography

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019): Bert: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics 1*, p. 4171–4186.

Ehrmann, Maud, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet (2022): Named Entity Recognition and Classification on Historical Documents: A Survey. URL: http://arxiv.org/abs/2109.11406.

Kamath, Uday, John Liu, and James Whitaker (2019): Deep Learning for NLP and Speech Recognition. Cham: Springer. URL: https://doi.org/10.1007/978-3-030-14596-5 [Access: 8th December 2021].

Labusch, Kai, Clemens Neudecker, and David Zellhöfer (2019): BERT for Named Entity Recognition in Contemporary and Historical German.

Nübling, Damaris, Antje Dammel, Janet Duke, and Renata Szczepaniak (2017): Historische Sprachwissenschaft des Deutschen. Eine Einführung in die Prinzipien des Sprachwandels. 5th revised and updated edition. Narr Francke Attempto: Tübingen.

Schmidt, Thomas, Katrin Dennerlein, and Christian Wolff (2021): Emotion Classification in German Plays with Transformer-based Language Models Pretrained on Historical and Contemporary Language.

Underwood, Ted (2019): Do Humanists need BERT? URL: https://tedunderwood.com/2019/07/15/do-humanists-need-bert/ [Access: 8th December 2021].

Zöllner, Jochen, Konrad Sperfeld, Christoph Wick, and Roger Labahn (2021): Optimizing small berts trained for german NER. arXiv. URL: https://arxiv.org/abs/2104.11559 [Access: 8th December 2021].

Notes

1. https://www.slm.uni-hamburg.de/germanistik/forschung/forschungsprojekte/dehmel-digital.html
2. https://www.bbaw.de/en/research/uwe-johnson-werkausgabe-the-complete-works-of-uwe-johnson
3. https://github.com/NEISSproject/tei_entity_enricher

Language Model Pre-Training for Historical English: Approaches and Evaluation

Fonteyn, Lauren
l.fonteyn@hum.leidenuniv.nl
University of Leiden, Netherlands

Manjavacas Arevalo, Enrique
enrique.manjavacas@gmail.com
University of Leiden, Netherlands

Machine-based exploration of culturally relevant datasets (e.g. newspapers, periodicals, correspondence or annals) often involves understanding and processing historical texts. In this context, technology dealing with historical text needs to tackle a set of specific issues. First, historical corpora typically contain not only a variety of registers and genres, but also an unstable language with grammar and semantics in constant change. Secondly, the lack of orthographic standards that characterizes European languages prior to the 18th centuries implies a further aspect of variation on the linguistic form over which automated approaches need to abstract. Finally, in contrast to contemporary corpora that have been born already digitally, historical corpora must be digitized. Despite ongoing efforts to advance OCR and HTR technology, errors in the digitization pipeline constitute the last barrier.

Current state-of-the-art approaches in Natural Language Processing (NLP) leverage the newly emerging pre-train-and-finetune paradigm, in which a large machine learning model is first trained in an unsupervised fashion on large

datasets, and then fine-tuned on labelled data in order to perform a particular task of interest. This paradigm relies on variants of so-called Language Models—e.g. ELMO (Peters et al. 2018), BERT (Devlin et al. 2019) or GPT (Radford et al., 2018)—that can maximally exploit contextual cues in order to generate vectorized representations of given input tokens (e.g. words). At first sight, this paradigm may seem unrealistic for dealing with the three aspects of historical text mentioned above, since the notoriously large datasets that are needed in order to successfully pre-train such models are absent in the case of historical text. However, on-going efforts towards producing historically pre-trained Language Models (Hosseini et al. 2021), which leverage large databases of available text, have started to highlight the potential of this approach even for historical languages.

With the goal of alleviating the data scarcity problem, previous research has leveraged contemporary Language Models (i.e. models pre-trained on contemporary data), using them as base models that are later fine-tuned on the target historical data in order to produce historically pre-trained models. While this approach may indeed help the model recognizing semantic relationships and linguistic patterns that are unchanged across time, it can also introduce an important bias in the vocabulary, which remains fixed to the vocabulary of the contemporary corpus.

In this work, we aim to cast light onto both the merit of the pre-train-and-fine-tune paradigm for historical data, as well as the relative advantage of the different pre-training approaches (e.g. pre-training from scratch vs. adapting a pre-trained model). We focus on a large span of historical English text (date range: 1450-1950), presenting, first, the steps towards a newly pre-trained historical BERT, known as MacBERTh, which is trained on approx. 3B tokens of historical English. Secondly, we discuss a thorough evaluation, with a noted emphasis on lexical semantics, in which the capabilities of the different models for processing historical text are put to test.

In order to do so, we elaborate a number of historically relevant benchmark tasks extracted from the Oxford English Dictionary, relying on the diachronic word sense classifications and the example quotations used for exemplifying the word senses.

In particular, we evaluate the different approaches on tasks that require systems to incorporate a model of word senses as well as a founded understanding of sentential semantics. In total, we present results for three different tasks, including Word Sense Disambiguation (WSD) from two different angles, Text Periodization and an ad-hoc Fill-In-The-Blank task that indirectly captures aspects of Natural Language Understanding.

Our evaluation highlights that, indeed, models originally pre-trained on contemporary English may also import too strong an inductive bias when they are later fine-tuned on historical English. In such a situation, pre-training from scratch on historical data may be a more robust strategy than adapting a pre-trained model.

With this work, we hope to assist Digital Humanities scholars willing to fine-tune and deploy NLP models on their own historical collections, as well as researchers who may be working on developing similar models and resources for other historical (and possibly non-European) languages.

Bibliography

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–86. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

Hosseini, Kasra, Kaspar Beelen, Giovanni Colavizza, and Mariona Coll Ardanuy. 2021. “Neural Language Models for Nineteenth-Century English.” *Journal of Open Humanities Data* 7 (September): 22. https://doi.org/10.5334/johd.48.

Peters, Matthew, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. “Deep Contextualized Word Representations.” In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–37. New Orleans, Louisiana: Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1202.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. “Improving Language Understanding by Generative Pre-Training.” *OpenAI*. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Reconstruction of cultural memory through digital storytelling: a case study of Shanghai Memory project

Fu, Yaming
ymfu@libnet.sh.cn
Shanghai Library/Institute of Scientific & Technical Information of Shanghai, Shanghai, China; School of

Information Management, Nanjing University, Nanjing, China

Mahony, Simon

SimonMahony@bnu.edu.cn
Research Centre for Digital Publishing and Digital Humanities, Beijing Normal University at Zhuhai, China

Liu, Wei

wliu@libnet.sh.cn
Shanghai Library/Institute of Scientific & Technical Information of Shanghai, Shanghai, China

The theory and practice of digital storytelling has been developing since the 1990s; firstly in media research, with a focus on audio-visual story creation using digital media (Lambert, 2018), and then extending into multiple fields such as public history (Burgess & Klaebe, 2009) and education (Robin, 2008), where there is a close relationship with human narrative. These fields discussed the possibilities for digital storytelling as they encountered the so-called "digital turn " which prompted the move from traditional storytelling research into the digital sphere and brought about epistemological as well as methodological shifts (Noiret, 2018).

Digital storytelling, understood here as a movement or method for creating, expressing, interpreting, and sharing stories and personal experiences using digital tools, has been viewed as a "democratization of culture." (Clarke & Adam, 2011) It draws attention from the mainstream to the marginalized, the minority, the overlooked and forgotten. Despite the discourse and practice of digital storytelling in education, history, and media research, its theory construction in DH and the practice in GLAMs is still at an exploratory stage.

Digital storytelling provides new opportunities for DH as both fields seek to encourage dialogue, make the world comprehensible, and discover new ways of interaction with the support of digital tools (Barber, 2016). Digital storytelling also serves as a bridge between cultural heritage and DH with "space and time as shared concepts," (Münster et al., 2019) essential dimensions in storytelling and other forms of narratives. A great potential exists for DH practitioners to employ GLAM collections to reconstruct knowledge and cultural heritage, discover hidden knowledge, and support knowledge creation through the lens of digital storytelling.

This presentation, drawing on extensive published literature and in-depth reflection, examines how digital storytelling is applied to encourage and facilitate cultural memory reconstruction as part of the Shanghai Memory project. [1] Relevant aspects focus on democratizing DH practice and the theory of cultural memory construction. *A Journey from Wukang Road* [2], a centrepiece of this project, associates the three dimensions of memory (the past), culture, and community as proposed by Assmann and Czaplicka (1995) to organize and construct the diverse collections pertaining to Wukang Road. Borrowing thinking from postcolonial studies around critical "re-reading " and "re-writing " of the colonial past, along with the continuing effect of memory (Ashcroft et al., 2002, p. 221), it recognizes and tells the holistic story of the past. Memory is achieved through knowledge organization and representation methods, including ontology design for people, places, time, events, architectures, etc.; resource description framework (RDF) to describe resources in a universal way and linked data to connect the entities. Culture is presented using both historical records from the library and contemporary reflections from the public. The community aspect engages citizens by having them upload photos and personal accounts of their memories and experiences of the road, adding to the underrepresented art forms housed in library collections (magazines, music recordings, photos, maps, and old movies), the places and people that constitute the history of Wukang Road.

Wukang Road is famous as the home of many celebrities and historic buildings going back to the colonial era, all having their own stories. Here with digital storytelling, sharing methodologies with oral and public history, we capture the voice of the common people so that the history and culture of Shanghai is democratized in the modern postcolonial era. While the buildings themselves are monuments to the formal history, the "road is the smallest unit of urban geography [and] another focus of urban memory is the space-time structure," (Xia et al., 2021, p. 849) which is why these stories fill the gaps over time and give voice to those usually unheard.

The project uses crowdsourcing method for the public to create digital stories, shifting them from the private to the public sphere, from "private forms of communication and translating them into contexts where they can potentially contribute to public culture." (Burgess & Klaebe, 2009, p. 155) Using digital storytelling in this way makes it an additional tool for researchers in public history and importantly "the recording of oral histories." (Earley-Spadoni, 2017, p. 97) The wider project identifies the material culture embedded in heritage objects and linked with sources makes "literature the historical witness for the material cultural heritage objects themselves." (Xia et al., 2021, p. 844) The personalized experience of citizens serve as an effective supplement to the formal literary accounts.

The Shanghai Memory project brings together many aspects of memory construction as part of a comprehensive programme of heritage management to more accurately reconstruct the history of the city (Xia et al., 2021). This

latest initiative to incorporate digital storytelling is the next phase to further democratize the practice and represent the unrepresented by presenting, creating, and sharing stories in relation to the past, current, and even the future of Shanghai city. This extension to an already established DH project adds significant value to the reconstruction of cultural memory and acts as a model for other memory projects in East Asia and beyond.

Bibliography

Ashcroft, B., Griffiths, G., & Tiffin, H. (2002). *The Empire Writes Back: Theory and Practice in Post-Colonial Literatures* (2nd ed.). Routledge.

Assmann, J., & Czaplicka, J. (1995). Collective Memory and Cultural Identity. *New German Critique*, *65*, 125–133.

Barber, J. F. (2016). Digital storytelling: New opportunities for humanities scholarship and pedagogy. *Cogent Arts and Humanities*, *3*(1).

Burgess, J., & Klaebe, H. (2009). Digital Storytelling as Participatory Public History in Australia. In J. Hartley & K. McWilliam (Eds.), *Story Circle: Digital Storytelling around the World* (pp. 155–166).

Clarke, R., & Adam, A. (2011). Digital storytelling in Australia: academic perspectives and reflections. *Arts and Humanities in Higher Education*, *11*(1–2), 157–176.

Earley-Spadoni, T. (2017). Spatial History, deep mapping and digital storytelling: archaeology's future imagined through an engagement with the Digital Humanities. *Journal of Archaeological Science*, *84*, 95–102.

Lambert, J. (2018). *Digital Storytelling: Capturing Lives, Creating Community* (5th ed.). Routledge.

Münster, S., Apollonio, F. I., Bell, P., Kuroczynski, P., Di Lenardo, I., Rinaudo, F., & Tamborrino, R. (2019). Digital Cultural Heritage meets Digital Humanities. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 813–820.

Noiret, S. (2018). Digital Public History. In D. Dean (Ed.), *A Companion to Public History* (1st ed., pp. 111–124).

Robin, B. R. (2008). Digital storytelling: A powerful technology tool for the 21st century classroom. *Theory into Practice*, *47*(3), 220–228.

Xia, C., Wang, L., & Liu, W. (2021). Shanghai memory as a digital humanities platform to rebuild the history of the city. *Digital Scholarship in the Humanities*, *36*(4), 841–857.

Notes

1. Shanghai Memory: http://memory.library.sh.cn
2. A Journey from Wukang Road: http://wkl.library.sh.cn

Building inclusivity into our digitization projects- a case study of digital collections in Mexico

Galina Russell, Isabel
igalina@unam.mx
Universidad Nacional Autónoma de México (UNAM), Mexico

Priani Saisó, Ernesto
epriani@gmail.com
Universidad Nacional Autónoma de México (UNAM), Mexico

Introduction

For several decades now there has been an emphasis on the importance of digitizing library collections, in particular bibliographic heritage materials. Although they are considered fundamentally useful and good for society in general, digitization projects are not neutral and they create new digital paradigms of our cultural memory (Thylstrup 2018) .

As part of our work for Oceanic Exchanges [1], we documented the creation of the HNDM, Mexico's national digital newspaper library. Using document analysis and interviews, we studied the complex interplay of technical, political and administrative decisions and the effect these have had on the resulting digital collection, its scope and usefulness (Galina Russell y Priani Saisó 2021; Galina Russell et al. 2020) . In particular, we noted that decisions regarding what should be digitized, and how it should be digitized (and then processed and presented), were not properly documented. This lack of context hampered our ability as researchers to properly interpret the results we were obtaining from text mining techniques. We found evidence that suggests that decisions regarding what was being digitized and how it would be presented were based mainly on previous decisions and were practical in nature, such as digitizing the microfilmed collection, rather than a critical inquiry of what this new digital collection would represent.

It was this that led us to develop a new research project which seeks to critically analyze digitization projects in Mexico, as we think that many of them have not conscientiously questioned broader aspects beyond digitization, such as the origins of the collections

themselves, the selection process for digitizing and how to (re)present the object digitally. In doing so, they simply transfer the existing biases and under and over representations of particular materials, to the new digital collections. Additionally, the physical artefact is simply reproduced in a digital format without providing access to the digital object as such, which in the case of textual collections, like the HNDM, would be the OCR text files.

The aim of this project is to critically examine how digitization projects have contemplated (if at all) these issues, and to propose new angles to thinking about digitization projects that can be more inclusive in representing traditionally marginalized populations, with a vision that goes beyond just visually representing the physical artefact and includes a digital representation that can be examined and used by computational methods.

Initial findings

Approaching digitization projects from this perspective is relatively new, and as such, we have worked on identifying relevant studies. There has been work done using critical race theory, feminist studies and postcolonial approximations to cultural digital collections in general, but we have found less work related to libraries and in particular digitization projects. Honma (2005) addresses issues related to race and LIS from a US perspective. Work has been done on post-colonial archives (Gauthereau 2018; Becerra-Liche 2017) , but we are particularly interested in digitization of library collections. There is also important work regarding libraries lack of neutrality (Galván 2015; Matienzo 2015; De Jesus 2014; Beatty 2014; Bourg 2015) . This work is relevant for our analysis but it is based in a different national context from ours in Mexico.

Latin America has a strong background in fields such as anthropology, sociology and political science, that examine the inequalities and exclusion which result from colonialism and post-colonialism, in particular related to the indigenous populations. Feminist studies have also addressed this in relation to the patriarchal system, gender violence and discrimination. In relation to LIS there is some work, in particular related to Open Access movement (Alperin, Fischman, y Marin 2015) and work related to information and indigenous populations in Mexico (Ramírez Velázquez y Figueroa Alcántara 2021) so far we have not identified work regarding digitization and libraries within this framework.

Future steps

This research project is currently in development. So far, we have established that there is indeed little work in this area, and we are proposing a particular framework from which to approach the next part of our work. We are now working on compiling a list of digitization projects in Mexico that have worked with cultural heritage textual materials. From here we shall select four case studies for in-depth work related to how the digitizing project was undertaken. We are also planning a workshop to receive further input from practitioners regarding these issues. Our final aim is to contribute towards creating more inclusive digitization projects.

Bibliography

Alperin, Juan Pablo, Gustavo Fischman, y Anabel Marin, eds. (2015). *Hecho en Latinoamérica: acceso abierto, revistas académicas e innovaciones regionales*. Primera edición en español. Brazil: FLACSO Brasil.

Beatty, Joshua. (2014). Locating Information Literacy within Institutional Oppression – In the Library with the Lead Pipe. https://www.inthelibrarywiththeleadpipe.org/2014/locating-information-literacy-within-institutional-oppression/.

Becerra-Liche, Sofía. (2017) Participatory and Post-Custodial Archives as Community Practice, *Educause Review*, 90-91.

Bourg, Chris. (2015). Never Neutral: Libraries, Technology, and Inclusion. *Feral Librarian* (blog). https://chrisbourg.wordpress.com/2015/01/28/never-neutral-libraries-technology-and-inclusion/.

De Jesus, Nina. (2014). Locating the Library in Institutional Oppression – In the Library with the Lead Pipe. https://www.inthelibrarywiththeleadpipe.org/2014/locating-the-library-in-institutional-oppression/.

Galina Russell, Isabel, y Ernesto Priani Saisó. (2021). Políticas de digitalización para la investigación. El caso de la HNDM y el proyecto Oceanic Exchanges, in *El estante digital*, 125-41. Santiago de Querétaro: Fondo Editorial de la Universidad Autónoma de Querétaro.

Galina Russell, Isabel, Laura Martínez Domínguez, Miriam Peña Pimentel, Ernesto Priani Saisó, y Rocío Castellanos Rueda. (2020). El uso de periódicos digitalizados como fuente para trabajos de investigación. *Relaciones Estudios de Historia y Sociedad* 43 (163).

Galván, Angela. (2015). Soliciting Performance, Hiding Bias: Whiteness and Librarianship – In the Library with the Lead Pipe, *In The Library With The Lead Pipe*. https://www.inthelibrarywiththeleadpipe.org/2015/soliciting-performance-hiding-bias-whiteness-and-librarianship/.

Gauthereau, Lorena. (2018). Post-Custodial Archives and Minority Collections. *Recovering the U.S. Hispanic Literary Heritage Blog* (blog). 7 de agosto de 2018. https://

recoveryprojectappblog.wordpress.com/2018/08/07/post-custodial-archives-and-minority-collections/.

Honma, Todd. 2005. Trippin' Over the Color Line: The Invisibility of Race in Library and Information Studies. *InterActions: UCLA Journal of Education and Information Studies* 1 (2). https://doi.org/10.5070/D412000540.

Matienzo, Mark A. 2015. To Hell With Good Intentions: Linked Data, Community, and the Power to Name. Available at https://matienzo.org/2016/to-hell-with-good-intentions/.

Ramírez Velázquez, César Augusto, y Hugo Alberto Figueroa Alcántara, eds. (2021). *La importancia de la información en las culturas originarias*. eSchola. Ciudad de México: FFyL, UNAM.

http://ru.atheneadigital.filos.unam.mx/jspui/bitstream/FFYL_UNAM/4615/1/La%20importancia%20de%20la%20informacio%CC%81n%20en%20las%20culturas%20originarias%20-%20EIPE.pdf.

Thylstrup, Nanna Bonde. (2018). *The politics of mass digitization*. Cambridge, MA: The MIT Press.

Notes

1. See Oceanic Exchanges- Tracing Global Information Networks in Historical Newspaper Repositories, 1840-1914: https://oceanicexchanges.org/

Having a Ball: A Linked Data Approach to Fancy Dress in Colonial Australia

Gatti, Tommy
u6044453@alumni.anu.edu.au
Australian National University, Australia

Nurmikko-Fuller, Terhi
terhi.nurmikko-fuller@anu.edu.au
Australian National University, Australia

Pickering, Paul
paul.pickering@anu.edu.au
Australian National University, Australia

Swift, Ben
ben.swift@anu.edu.au
Australian National University, Australia

Introduction

The *Lord Mayor's Costume Balls in Sydney in 1857 and 1879* (LMB) is a prototype that focuses on a single page in a vast archive: a list of attendees at a fancy-dress costume ball, hosted by the Lord Mayor of Sydney in the British Colony of New South Wales in 1857, and published in the *Sydney Morning Herald*, the colony's leading newspaper. The tabular dataset is structurally simple, containing the names, titles, and the fancy-dress costumes worn by 994 invited guests: it is captured in 6,347 RDF triples. A prosopographical analysis of this list provides insight into the vicissitudes of Sydney's socio-political composition.

The overarching archive and the LMB ontology have been described elsewhere (Nurmikko-Fuller and Pickering, 2021): the latter is simple, but fit-for-purpose to prove the suitability of Linked Data (LD) for enriching scholarship into the origins of Australia's modern politics.

Background

The value of connecting complementary data across disparate datasets has been a feature of the study of Australian history for decades (e.g. Pope and Withers, 1993; Holman, et al., 1999; Moses, 2004). Outside of Australia, LD has been applied very successfully to historical investigations (Rantala et al., 2021; Schmidt and Eggert, 2019; Kaplan et al., 2021; Meroño-Peñuela et al., 2015; Dijkshoorn et al., 2014; de Boer, 2015). Although the potential is clear, few projects have successfully combined the two. Part of the problem is that SPARQL endpoints can be cryptic, lack helpful error messages or executable suggestions, and require prior knowledge of the syntax (Ngonga et al., 2013)., Past attempts to solve this problem (Russell et al., 2008; Lohmann et al., 2016; Haag et al., 2015; Ochieng, 2020; Pradel, 2014; Yang et al., 2018) have not explicitly focused on Humanities data.

In recognition of this dichotomy, we developed a bespoke user interface that enables researchers with little or no prior experience of SPARQL to engage with the LMB's knowledge graph. We have dubbed this the LMB SPARQL Explorer.

Bespoke UI

The SPARQL Explorer (Figure 1) consists of the Suggestion Generator (SG); Canvas, and Graph-to-SPARQL compiler. It is a single-page web application made using React, hosted on an Apache web server, utilising an Blazegraph-mandated API.

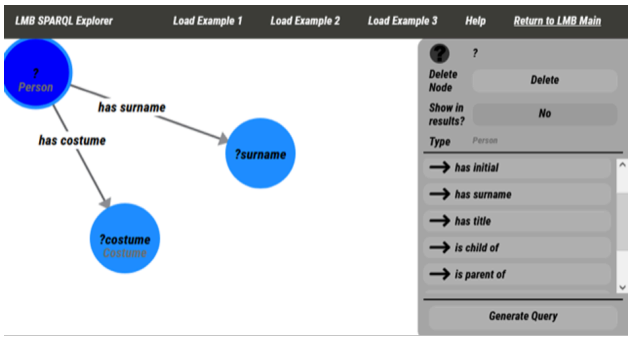


Figure 1:
The LMB SPARQL Explorer

A deliberate design feature was ontology-agnosticism: it should work equally well with any ontology. The SG enforces syntactic integrity through querying the triplestore for (any) ontologies; when found, they are cached locally and deconstructed into constituent Classes and properties whilst retaining metadata (e.g. comments, notes). Semantically useful queries are generated through the detection of each domain and range. The Canvas converts the query into a set of .SVG elements and displays it as a graph stored in two arrays, one for the state of nodes and the other for edges. A 1-1 mapping between the graph format and query language syntax ensures that every valid graph is a valid SPARQL query.

Example Query

Figure 2 illustrates the function of the SPARQL Explorer. The graph has been created by the user dragging suggested Classes and properties from the grey panel on the far right into the white space. Behind the scenes, a SPARQL query is generated.

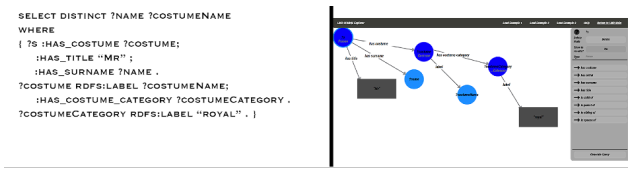


Figure 2:
Query and its representation in the LMB SPARQL Explorer

This query (left, Figure 2) produces a result set of 32 individuals who have the title of "Mr" and a costume categorised as "royal". It is small enough to enable domain experts to use tacit prior knowledge to infer further knowledge about who among them were royalists, dressed in homage to their monarchical ideals, and how many, in turn, donned royal garbs as a form of satire.

Preliminary Evaluation

The SPARQL Explorer was preliminarily tested with bespoke (JazzCats ontology, the LMB ontology) and ISO-standard ontologies (CIDOC-CRM and FOAF). For three out of the four, every possible path was representable: there was 100% coverage. The JazzCats ontology had 84% coverage: properties and Classes that had blank nodes as domains and/or ranges were inaccessible.

Conclusion

We have reported on a prototype web-based interface that can leverage any ontological structure to deliver syntactic validity and semantically useful queries over RDF without the need to learn SPARQL explicitly. Our preliminary testing has shown that a conceptual mapping between a visual query language and SPARQL is possible. What has been achieved is portentous: a pointer to a way forward for domain experts to seek richer answers by asking more complex questions of their (Linked) data.

Bibliography

de Boer, V. (2015). *Linked Data for Digital History. Semantic Web for Scientific Heritage, Proceedings of the Twelfth Extended Semantic Web Conference*, Portoroz, Slovenia, March 2015.

Dijkshoorn, C., Aroyo, L., Schreiber, G., Wielemaker, J., and Jongma, L. (2014). *Using linked data to diversify search results: a case study in cultural heritage, Proceedings of the Nineteenth International Conference on Knowledge Engineering and Knowledge Management*, Linkoping, Sweden, November 2014.

Haag, F., Lohmann, S., Siek, S., and Ertl, T. (2015). *QueryVOWL: A visual query notation for linked data', Proceedings of the Twelfth Extended Semantic Web Conference*, Portoroz, Slovenia, March 2015.

Holman, C., D'Arcy, J., Bass, J., Rouse, I.L., and Hobbs, M.S.T. (1999). Population -based linkage of health records in Western Australia: development of a health services research linked database, *Australian and New Zealand Journal of Public Health,* 23,5: 453-9

Kaplan, F., Oliveira, S. A., Clematide, S., Ehrmann, M., & Barman, R. (2021). Combining visual and textual features for semantic segmentation of historical newspapers, *Journal of Data Mining & Digital Humanities, HistoInformatics*, 19 January, 2021.

Lohmann, S., Negru, S., Haag, F., and Ertl, T. (2016). Visualizing ontologies with VOWL, *Semantic Web,* 7, 4: 399-419.

Meroño-Peñuela, A., Ashkpour, A., Van Erp, M. Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., and Van Harmelen, F. (2015). Semantic technologies for historical research: A survey, *Semantic Web*, 6, 6: 539-64.

Moses, A. D. (ed), (2004). *Genocide and settler society: Frontier violence and stolen indigenous children in Australian history*, vol. 6. Berghahn Books, New York.

Ngonga, N., Cyrille, A., Bühmann, L., Unger, C., Lehmann, J. and Gerber, D. (2013). *Sorry, I don't speak SPARQL: translating SPARQL queries into natural language, Proceedings of the 22nd international conference on World Wide Web*, Rio de Janeiro, Brazil, May 2013.

Nurmikko-Fuller, T., and Pickering, P. (2021). *Reductio ad absurdum?: From Analogue Hypertext to Digital Humanities, Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, Dublin, Ireland, September 2021.

Ochieng, P. (2020) 'PAROT: Translating natural language to SPARQL', *Expert Systems with Applications: X* 5, Article100024.

Pope, D. and Withers, G. (1993) Do migrants rob jobs? Lessons of Australian history, 1861–1991, *The Journal of Economic History*, 53, 4: 719-742.

Pradel, C.,Haemmerlé, O., and Hernandez, N. (2014) *'Swip: A natural language to sparql interface implemented with sparql', Proceedings of the Fourteenth International Conference on Conceptual Structures*, Iasi, Romania, July 2014.

Rantala, H., Ikkala, E., Jokipii, I., & Hyvönen, E. (2021) WarVictimSampo 1914–1922: a National War Memorial on the Semantic Web for Digital Humanities Research and Applications. *Journal on Computing and Cultural Heritage*, 15, 1: 1-18.

Russell, A., Smart, P., Braines, D., and Shadbolt, N. (2008) 'NITELIGHT: A graphical tool for semantic query construction', *Proceedings of the Conference on Human Factors in Computing Systems,* Florence, Italy, April 2008.

Schmidt, D., and Eggert, P. (2019) The Charles Harpur Critical Archive. *International Journal of Digital Humanities*, 1, 2: 279-288 **.**

Yang, C., Wang, X., Xu, Q., and Li, W. (2018) *SPARQLVis: an interactive visualization tool for knowledge graphs'*, *Proceedings of the Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*, Macau, China, July 2018.

Crowdsourced Distributed Open Collaborative Courses (DOCC) for inclusive, self-regulated learning: A study on OERs in India

Ghosh, Sharanya
ghosh.5@iitj.ac.in
Indian Institute of Technology, Jodhpur, India

Das, Rajarshi
das.16@iitj.ac.in
Indian Institute of Technology, Jodhpur, India

Our inquiry delves into the pedagogical potential of openly accessible educational and informational resources, in the form of Distributed Open Collaborative Courses or DOCC, as they might be introduced in the Indian higher education context or the larger Asian contexts, offering the scope of incorporating diverse ideas and materials from indigenous resources as we promote a crowdsourced and inclusive learning environment. Open Educational Resources are the net result of significant technology and policy shifts in the field of education. Massive Open Online Courses or MOOCs are one of the most prominent examples of OERs that engage learners from across the globe via the internet. However, David Wiley, a major voice in open education, has questioned the genuine openness of MOOCs in his historical overview of the sector. Similarly, Karen Head has highlighted some of these concerns regarding MOOCs, such as the possibility of exclusivity in such courses and the privileging of certain dominating voices, particularly those from specific parts of the world and socioeconomic status (SES) levels. **FemTechNet** is a breakthrough initiative that grew out of a network of scholars, artists, and students interested in the convergence of technology, science, and feminism. They work together to provide everyday course experiences and content within individually directed courses in a MOOC derivative known as a Distributed Open Collaborative Course or DOCC. The **GE Academy** is a DOCC built around 7 mini-courses that represent the main thematic areas as suggested by the European Commission to build a Gender Equality Plan (GE Academy, 2019).

India is a country of rich diversity and deep-rooted divisions. As we rethink higher education pedagogy in the Indian context to offer inclusive, participatory, and open-access education to a larger learner community, many of whom are victims of the great digital divide in the country, DOCCs come to offer some solutions. One important observation made by Aparicio et al. (2015) is that "DOCC... differentiates from MOOC in its focus on the pedagogic engagement of all actors, underlining, on the one hand, the invisible work of teachers, and on the other the collective intelligence of scholars." The use of DOCC as a democratising and sensitising tool for issues like gender, caste, class, race etc., has already been established by FemTechNet and GE Academy. Our proposal here is to visualise the DOCC as a space allowing its stakeholders to take part in course designing to create crowdsourced content, accessible through an open platform that engages itself with everyday issues and experiences emanating from the fault lines of caste, gender, and marginalisation

of the indigenous population, which also becomes a new pedagogical model for online education. Such a bottom-up approach may result in greater learner involvement in researching, exploring, reflecting, and deciding what and how they want to learn. This learner-oriented approach is so far missing in the MOOCs, such as those offered by SWAYAM-NPTEL in India. Open learning is also about open access, which may be facilitated by the digital revolution's strategic changes in media infrastructure. The flexible and collaborative format that DOCCs offer may help underprivileged learners learn at their own pace in a culturally sensitive digital space while engaging in curating the platform/course contents and metadata.

In our presentation, we will look at the two existing DOCCs mentioned above as we attempt a comparative analysis of the relevance of DOCC as an OER for Indian higher education, focusing primarily on our proposed model of a crowdsourced, learner-centric, open-access platform for training into some socially relevant issues. We have tried to develop some critical questions from detailed surveying of the SWAYAM-NPTEL MOOCs and the extant literature. By reviewing existing course contents and platforms, we now intend to test the efficacy of these existing coursewares in the South Asian context, particularly in India. Moreover, we have also been reflecting upon the kind of courses, their content, and the metadata for such courses. We contend that a top-down approach in designing such need-based courseware will only end in futility. Instead, our DOCC prototype will include features such as qualitative questions for users to arrive at pertinent keywords, crowdsourced course design, and multimodal, multilingual interface to address issues of digital literacy, digital divide, learner-centred pedagogy, and the more technical issues such as the limitations of the existing subject heading norms, such as the Library of Congress Subject Heading (LCSH) List, in the context of DOCCs.

Bibliography

Aparicio M., Bacao F., and Oliveira T. (2016). An e-Learning Theoretical Framework. In Altanay, F, Cagiltay, K., Gemini, M., and Altanay, Z. (eds.), *Journal of Educational Technology & Society*, *19* (1):292–307, http://www.jstor.org/stable/jeductechsoci.19.1.292 (accessed: 22 March 2022).

Bonk, C. J., Lee, M. M., Reeves, T.C., and Reynolds, T.H. (eds.) (2015). *MOOCs and Open Education around the World*, New York:Taylor & Francis.

Edwards, J. C. (2015). Wiki Women: Bringing Women Into Wikipedia through Activism and Pedagogy. *The History Teacher*, *48* (3), 409–436. http://www.jstor.org/stable/24810523

FemTechNet, (2013). *Transforming Higher Education with Distributed Open Collaborative Courses (DOCCs): Feminist Pedagogies and Networked Learning,* FemTechNet White Paper Committee, https://www.femtechnet.org/about/white-paper/ (accessed: 19 April 2022).

GE Academy, (2019) **.** *Gender Equality Training Materials,* GE Academy, https://ge-academy.eu (accessed: 19 April 2022).

Head, K. J. (2017). Talking Business in Higher Education: *Disrupt This! - Moocs and the Promises of Technology*. Lebanon: University Press Of New England.

Smith, M. N. (2014). Frozen Social Relations and Time for a Thaw: Visibility, Exclusions, and Considerations for Postcolonial Digital Archives. *Journal of Victorian Culture*, *19 (* 3): 403–410, https://doi.org/10.1080/13555502.2014.947189 (accessed 20 March 2022).

Wiley, D. (2018). Open Educational Resources: *Foundations of Learning and Instructional Design Technology.* EdTech Books, https://lidtfoundations.pressbooks.com/chapter/open-educational-resources/ (accessed: 30 March 2022).

Sentiment lexicons or BERT? A comparison of sentiment analysis approaches and their performance.

Grisot, Giulia
giulia.grisot@uni-bielefeld.de
Bielefeld University, Germany

Rebora, Simone
simone.rebora@uni-bielefeld.de
Bielefeld University, Germany

Herrmann, Berenike
berenike.herrmann@uni-bielefeld.de
Bielefeld University, Germany

Abstract

With the development of new powerful technologies for computational data analysis, the opportunities for – and interest in – the detection of sentiments and opinion in texts have grown considerably (Liu 2015). Because of the vast amount of material available online, these investigations have focused mostly on textual material gathered from

social media, making use of traditional corpus linguistic approaches as well as deep learning tools.

Sophisticated sentiment analysis (SA) of literary texts, especially in languages other than English, is still in its infancy (Kim and Klinger 2019). This has depended partly on the limited amount of digital texts available, partly the complex structure of literary texts, and finally on methodological challenges, with skills needed that seldom form part of the training of literary scholars.

Emotions are however central to the experience of literary narrative (Oatley 2012; Hogan 2016), and recent advances in their systematic, quantitative analysis have been made within computational literary studies (Jockers 2017, Burghardt et al 2019). Yet, such investigations have mostly relied on existing lists of words associated with sentiment and emotion values, the so-called *sentiment lexicons*. While these remain conventional and useful tools, they can only provide limited insights to the representation of emotions in texts.

Using a lexicon-based method, we have previously investigated emotions and sentiments in relation to the representation of landscape in Swiss literature, looking in particular at the differences between the *rural* and *urban* spaces portrayed in a corpus of Swiss novels written in German (see Grisot, G., Herrmann, J.B. (in preparation).

The present paper takes a step forward, using manual annotation and advanced machine learning methods to train a fine-tuned model to recognise *valence* and *arousal* on a historical corpus. Our goals are higher levels of lexical coverage and validity when compared to our prior results obtained with sentiment lexicons.

We describe here the current state of our method to detect sentiment using deep learning approaches. Using a language model trained on a large corpus (3000+) of German literary texts spanning from 1800 to 1950 (Fischer and Strötgen 2017), we make use of BERT word embeddings and manually annotated sentences to recognise sentiment.

500+ sentences were taken from three Swiss-German novels - one of which children's fiction - and annotated for *valence* - understood here as the degree of 'positivity' of the detected sentiment - and *arousal* - its 'intensity' or 'degree of activation' - by two trained student assistants using the same instructions. Currently, intra-class correlation coefficient (ICC) between manual annotators calculated on these sentences is 0.721 for valence and 0.606 for arousal.

Scores for individual texts indicated preliminary evidence of a genre effect, with higher ICCs for the children's novel (valence 0.86; arousal 0.78 for 149 sentences) as compared to the other two, more complex, realist novels (valence 0.78, 0.62 arousal for 182 sentences; 0.51 for valence, 0.41 arousal for 198 sentences).

The annotated sample was used to train a deep learning classifier, using linear regression to finetune the Literary German BERT model (https://huggingface.co/severinsimmler/literary-german-bert), which reached Pearson's r scores of 0.53 for valence and 0.63 for arousal. These scores are very promising, suggesting the possibility - provided more training data - of a full automation of the annotation task on our domain of historical literary texts.

We are currently appending the annotation and at the time of the conference shall be able to update the results on a broader data base.

While potentially taking automatic SA of German literary texts to a new level, our study also allows evaluating the performance of lexicon-based in direct comparison with deep learning SA approaches, thus allowing to gauge the validity of different SA methods on a data-driven basis. This approach also raises questions concerning the effect of genre on the ease and validity of manual sentiment annotations.

Bibliography

Burghardt, M., Wolff, C., & Schmidt, T. (2019, January 1). Toward multimodal sentiment analysis of historic plays: A case study with text and audio for Lessing's Emilia Galotti. 4th Conference of the Association Digital Humanities in the Nordic Countries.

Fischer, F., & Strötgen, J. (2017). Corpus of German-Language Fiction (txt). https://doi.org/10.6084/m9.figshare.4524680.v1

Grisot, G. and Herrmann, B. J. (in preparation) Examining the representation of landscape and its emotional value in German-Swiss fiction around 1900

Hogan, P. C. (2016). Affect Studies. Oxford University Press. https://doi.org/10.1093/acrefore/9780190201098.013.105

Jockers, M. (2017). Extracts sentiment and sentiment-derived plot arcs from text. R Package "Syuzhet.

Kim, E., & Klinger, R. (2019). A survey on sentiment and emotion analysis for computational literary studies. Zeitschrift Für Digitale Geisteswissenschaften. https://doi.org/10.17175/2019_008

Liu, B. (2015). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. New York, NY: Cambridge University Press.

Oatley, K. (2012). The Passionate Muse: Exploring Emotion in Stories. New York: Oxford University Press.

R Core Team. (2021). R: A language and environment for statistical computing. https://www.r-project.org/

Event annotation for literary corpora analysis

Grunspan, Claude
claude.grunspan@sorbonne-nouvelle.fr
Lattice (ENS-PSL & CNRS & Université Sorbonne nouvelle)

Mélanie, Frédérique
frederique.melanie@ens.psl.eu
Lattice (ENS-PSL & CNRS & Université Sorbonne nouvelle)

Barré, Jean
jean.barre@chartes.psl.eu
Lattice (ENS-PSL & CNRS & Université Sorbonne nouvelle)

Chardon, Laurette
Laurette.chardon@unicaen.fr
Crisco (Université de Caen)

Galleron, Ioana
ioana.galleron@sorbonne-nouvelle.fr
Lattice (ENS-PSL & CNRS & Université Sorbonne nouvelle)

Naguib, Marco
marco.naguib@hotmail.com
Lattice (ENS-PSL & CNRS & Université Sorbonne nouvelle)

Plancq, Clément
clement.plancq@univ-tours.fr
MSH Val de Loire (Université de Tours)

Seminck, Olga
olga.seminck@cri-paris.org
Lattice (ENS-PSL & CNRS & Université Sorbonne nouvelle)

Poibeau, Thierry
thierry.poibeau@ens.fr
Lattice (ENS-PSL & CNRS & Université Sorbonne nouvelle)

Studying large corpora in the literature domain, especially novels, mean that new tools are needed in order to address narratological questions at scale. A large body of research has specific developed techniques for the task, giving birth to the field known as distant reading (as opposed to close reading, by a human being), (Moretti, 2013). In this paper, we present a series of tools providing the basis for the large-scale and comprehensive annotation of French novels through the adaptation of the BookNLP project (Bamman et al. 2014) to French. We present the different kinds of annotation provided and then address specific issues concerning the annotation of events (Vauth *et al.*, 2021).

Event annotation within the BookNLP project

The BookNLP framework (Bamman et al. 2014) is one of these software ensembles integrating various modules (entity recognition [1], coreference [2], event and quotation analysis [3]) that can be applied to large collections of text. The initial BookNLP contained tools for English only, and a new project is now extending the range of languages covered. We are on our side developing the same kind of modules for French.

Natural language processing is now almost exclusively based on machine learning techniques, which means most of the effort required to develop this kind of tools lies in text annotation. For French, we have annotated 20 extracts of French novels from the 19 th and 20 th century. We build on the Democrat project [4], whose aim was to annotate a large corpus of French texts (from different historical periods and different genres) with coreference information. We selected the texts corresponding to our criteria (copyright free texts from novels from the 19 and early 20 th century), hence our 20 extracts (for a total of 184.000 words).

The task first consisted in annotating entities following the BookNLP guidelines and mapping the initial Democrat coreference annotations to BookNLP. We then focused on event annotation, as this is one of the key features for distinguishing between author styles, but also for identifying specific episodes in a story, such as the fortune changes of the main characters, or the climax of a story arc.

However, we discovered that annotating events is slightly more difficult than annotating entities. In BookNLP (Sims et al., 2019, Bamman et al., 2019 and 2020), the definition of the notion of event is as follows: “The event layer identifies events with asserted realis (depicted as actually taking place, with specific participants at a specific time) – as opposed to events with other epistemic modalities (hypotheticals, future events, extradiegetic summaries by the narrator)”. The definition entails that verbs with a

negation or with a modal are not annotated, for example, and only conjugated form of the verbs are annotated.

The necessity to integrate modals and negation in the annotation scheme

We chose to annotate all kinds of events, without the initial limitations imposed in BookNLP. The first example presents three sentences with approximately the same meaning. If we leave apart the conjugated verb in the main clause, all the sentences include another clause, with a conjugated verb in the first sentence (1a), with an infinitive in the second (1b), and with a participle in the third one (1c).

1a. Après qu'il a mangé, il s'en est allé.

1b. Après avoir mangé, il s'en est allé.

1c. Ayant mangé, il s'en est allé.

1d. *After he had eaten, he left.*

1a – 1c have roughly the same meaning and should thus be annotated with two events, independently of the form of the verb in the subordinate clause.

Negation is more complex, as *generally* a negation means that no event has occurred. But this is not always the case and examples like 2a can be found:

2a. Il ne put retenir ses larmes.

2b. *He could not hold back his tears.*

which roughly means that the character cried. In an example like this one, there is definitely an action so in our opinion it should be annotated as such. Here our choices differ slightly from the ones in the original BookNLP project.

All annotations were carried out after multiple rounds of discussions and the creation of a set of annotation guidelines heavily dependent on the initial BookNLP annotation scheme for events (but including the differences highlighted in this section). The total dataset comprises 14,305 events among 184,000 tokens in the 20 books in our corpus.

The annotated corpus as well as our guidelines are freely available on GitHub. A collection of computer programs makes it possible to go from our annotation to something close to the original BookNLP scheme by excluding from the corpus examples with a negation or a modal. The next steps will consist in evaluating the robustness of the developed solution and its ability to provide useful information for actual literary studies.

Bibliography

Bamman D., Underwood T. and Smith N. (2014). A Bayesian Mixed Effects Model of Literary Character, *Proceedings of the conference of the Association for Computational Linguistics* (ACL), Baltimore, USA, June 2014.

Bamman D., Popat S. and Shen S. (2019). An Annotated Dataset of Literary Entities. *Proceedings of the conference of the North American Association for Computational Linguistics* (NAACL), Minneapolis, USA, June 2019.

Bamman D., Lewke O. and Mansoor A. (2020). An Annotated Dataset of Coreference in English Literature. *Proceedings of the Language and Resource Evaluation Conference* (LREC), Marseille, France, May 2020.

Landragin F. (2021), Le corpus DEMOCRAT et son exploitation. *Langages* nº 224 (4/2021), pp. 11-24,

Moretti F. (2013), *Distant Reading*, Verso Books, London.

Sims M., Park J.H. and Bamman D. (2019). Literary Event Detection. Proceedings of the Conference of the Association for Computational Linguistics. Florence, Italy, July 2019.

Vauth M., Hatzel H.O., Gius E. and Biemann C. (2021). Automated Event Annotation in Literary Texts. CHR 2021: Computational Humanities Research Conference, Amsterdam, The Netherlands, November 2021.

Notes

1. Person names, location names, etc.
2. A coreference occurs when two or more expressions refer to the same person or thing, like in *Joe Biden* $_i$ *said… He* $_i$ *was… The president* $_i$ *appeared to be…*
3. Roughly, who (the source) said what (the quotation).
4. https://www.ortolang.fr/market/corpora/democrat/3

Shared Tasks in Computational Literary Studies: Guideline Creation, Annotation and Text Generation for the Analysis of Narrative Levels

Guhr, Svenja Simone
svenja.guhr@tu-darmstadt.de
Technical University of Darmstadt, Germany

Reiter, Nils
nils.reiter@uni-koeln.de
University of Cologne

Zarrieß, Sina
sina.zarriess@uni-bielefeld.de
University of Bielefeld

Gius, Evelyn
evelyn.gius@tu-darmstadt.de
Technical University of Darmstadt, Germany

Introduction

With this short presentation we would like to give an update on our ongoing shared tasks for text annotation. In natural language processing, shared tasks are well established and highly productive frameworks for bringing together researchers. Organizers define a research task and the conditions for a competition-like setting in which others can submit their approaches to the task. We believe that shared tasks are a productive way of collaboration for the digital humanities and thus we hope that our presentation encourages others to organize shared tasks. We started to design our shared tasks on the identification of narrative levels in 2017. Here, we report on the successful conclusion of our first shared task on guideline creation and give an outlook on the subsequent task of automated annotation.

The goal of our shared task is the annotation of narrative levels. Automated segmentation of prose texts into meaningful units remains challenging, but would enable more robust and meaningful subsequent processing tasks, such as authorship determination, computation of topics, sentiment analysis, or procedures for determining semantic similarities, because they can be better connected to literary studies theories and assumptions (e.g., the characteristics property of an author might be a certain rhythm in their texts). In addition, any content-related analysis of narrative texts (e.g., about the characters or the plot) greatly benefit from knowledge of the level structure. Narrative levels are segments of prose texts determined by a largely stable expression of their fictional world in terms of space, time, and events, as well as the mediation of the events by a narrator.

First Shared Task on the Systematic Annotation of Narrative Levels

In an initial shared task, which we had announced in a presentation at DH2017 (Reiter et al., 2017), eight teams of linguists, computational linguists and literary scholars developed guidelines for the annotation of narrative levels. These reflect the diverse, heterogeneous theories of narrative levels. (Cf. Gius et al. (2019) for the setting and the guidelines of the initial round and Gius et al. (2021) for the guidelines of the second round and a final overall reflection on the shared task.)

Each of the submitted guidelines had strengths and weaknesses that emerged in the evaluation (Gius et al., 2019; Gius et al., 2021).

Second Shared Task on the Automatic Annotation of Literary Texts and Generation of Training Data

A second shared task with a different focus is currently in preparation. It uses the level-annotated texts from the first shared task as the basis for creating a training and test set for developing automated level recognition methods. In addition to automating level recognition, it also addresses the automated generation of training data, since manual annotation is expensive and time-consuming.

The second shared task will consist of two connected but independently solvable tracks, and it will focus on the English language.

Track One: Automating Generation of Level-Annotated Training Data

Track participants address the task to develop procedures for the automatic generation or synthesis of texts that are organized in terms of different narrative levels. We expect participants to experiment with different rule-, template- and language-model/AI-based techniques to assemble texts that display typical patterns of narrative level changes. The resulting texts will be used as training data for a BERT-based baseline system that automatically learns to detect narrative levels. Hence, the objective of narrative level generation is not primarily to produce high-quality, aesthetically pleasing text, but rather to obtain data of sufficient quantity and quality for supporting machine learning techniques in narrative level detection.

Track Two: Automatic Detection of Narrative Levels in Literary Prose

The second track builds on experiments on automatic detection of narrative level boundaries using a pre-trained BERT-model to detect artificially generated level boundaries that we conducted in 2021 (Reiter et al., 2022).

Participants will use annotated training data (some of which will be generated in the first track) to develop systems for automatic detection of narrative levels in texts.

Evaluation

The evaluation of the first track will be done via the performance of the BERT-based recognition model, using the generated data as training data. Thus, the generated texts are measured on whether they contain well-recognizable narrative levels that are useful for training. The second track will be evaluated using manually annotated gold data from the first shared task in 2020.

Bibliography

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. DOI 10.48550/arXiv.1810.04805.

Gius, E., Willand, M., Reiter, N. (2021). On Organizing a Shared Task for the Digital Humanities – Conclusions and Future Paths. *Journal of Cultural Analytics*, 6(4). DOI 10.22148/001c.30697.

Gius, E., Reiter, N., Willand, M. (2019). A Shared Task for the Digital Humanities Chapter 2: Evaluating Annotation Guidelines. *Journal of Cultural Analytics*, 4(3). DOI 10.22148/16.049.

Reiter, N., Sieker, J., Guhr, S., Gius, E., Zarrieß, S. (2022). Exploring Text Recombination for Automatic Narrative Level Detection. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*.

Reiter, N., Gius, E. Strötgen, J., Willand, M. (2017). A Shared Task for a Shared Goal: Systematic Annotation of Literary Texts. *Book of Abstracts*. DH 2017. Montreal, Canada: ADHO. https://dh2017.adho.org/abstracts/192/192.pdf.

Reiter, N., Willand, M., Gius, E. (2019). A Shared Task for the Digital Humanities Chapter 1: Introduction to Annotation, Narrative Levels and Shared Tasks. *Journal of Cultural Analytics*, 4(3), DOI 10.22148/16.048.

Ryan, M.-L. (1991). Possible Worlds, *Artificial Intelligence, and Narrative Theory*. Bloomington: Indiana University Press.

Towards a prosopographical ecosystem: modelling, design, and implementation issues

Hadden, Richard William James
richard.hadden@oeaw.ac.at
Austrian Academy of Sciences, Austria

Schlögl, Matthias
matthias.schloegl@oeaw.ac.at
Austrian Academy of Sciences, Austria

Vogeler, Georg
georg.vogeler@uni-graz.ac.at
Austrian Academy of Sciences, Austria; University of Graz, Austria

The International Prosopography Interchange Format

This presentation intends to describe ongoing work at the Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH) on the International Prosopographical Interchange Format (IPIF). It presents IPIF's design, and explores various conceptual and technical challenges arising from its implementation.

IPIF is an API definition and data model enabling the sharing and querying of prosopographical data. The original IPIF paper (Vogeler et al. 2019) recognises the power of semantic web tools (RDF, OWL, SPARQL), but also highlights their shortcomings for an interoperable format, chiefly the absence of a standard data model and the complexity of SPARQL. (See Bradley 2020.) Such difficulties circumscribe uses such as 'light-weight' data access and querying (as opposed to complex data processing). Accordingly, IPIF opts for a simple RESTful API.

In addition to a reference implementation, Papilotte (Vasold 2019), IPIF is intended to be implemented on top of existing projects, allowing access to data in a standardised format and providing the facilities for federated queries and as a data discovery tool. (Vogeler et al. 2019)

The Data Model and API

To achieve maximum interoperability, IPIF adopts a version of the 'factoid model' (Bradley 2005). This model separates *statements made about a person* from the abstract idea of a *person*, instead attributing statements to a researcher's interpretation of a historical *source* (Figure 1). This enables contradictory statements, made by distinct sources, to be recorded. As Baillie (2021) argues, this is not always desirable, some historical sources being more trustworthy than others; nevertheless, it is a requirement for an open ecosystem without any single 'guiding hand' that contradiction be permitted.

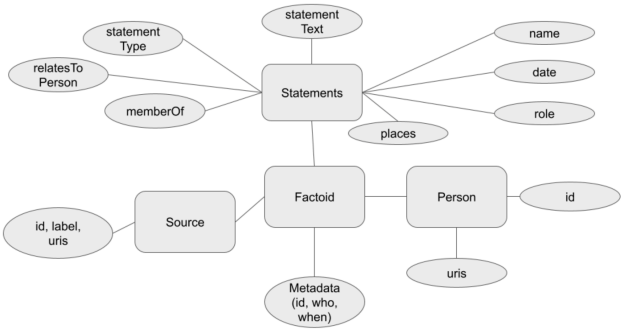


FIGURE 1:
THE IPIF DATA MODEL

The API — described in OpenAPI — is a RESTful interface, allowing direct access to the four IPIF entity types (Factoid, Source, Person, Statement), and querying of statements through graph-like traversals (Figure 2). It returns JSON-LD for each entity, comprising content and/or metadata, and embeds references to related entities.

The choice of Statement fields represents a highly pragmatic decision. These fields, allowing a label and a URI, were chosen to match the most obvious use-cases in prosopographical data; the *statementType* field allows arbitrary extension beyond these standard statement types.

```
/persons/tangl_k  (get a person by ID or by URI)
/persons/http://apis.acdh.ac.at/entity/1408/

/statements/tangl_k_name  (get statement by ID)
/factoids/factoid_27527  (get factoid by ID)
/sources/original_source_594  (get statement by ID)

/sources?personId=tangl_k (all sources that are related, via some factoid, to the
person with ID 'tangl_k')

/statements?name=Tangl  (statements assigning the name 'Tangl')

/person?role=banker&memberOf=CreditSuisse (persons connected via a factoid
to a statement that assigns the role 'banker' in the organisation 'CreditSuisse')
```

FIGURE 2:
API EXAMPLES

Implementations, issues and works-in-progress

Since the original definition of the standard, IPIF endpoints have been implemented on top of several existing frameworks, including the ACDH-CH's APIS platform (a Django- based platform for prosopographical projects) and as an eXist-DB module for serving TEI-XML personography data (typically from digital scholarly editions). IPIF client libraries for Python and JavaScript, allowing federated queries and data aggregation across several endpoints, have also been developed. Building these tools has afforded a wealth of practical experience, highlighting the strengths of the format and the difficulties involved in its implementation. This has led to several pragmatic decisions regarding the data model, API and the semantics of querying.

- A *label* field was introduced for Persons, allowing use cases such as autocompletes. The theoretically correct modelling — as a 'naming statement' — required too many additional requests to retrieve the appropriate information.
- Persons and Sources allow multiple URIs (interpreted as owl:sameAs). Strictly, these should be Statements (i.e. non-definitive assertion that one Person or Source is the same as another: see Zedlitz 2009); but reconciling data from multiple endpoints is considerably easier when this information is a 'meta' field of a Person entity.
- IPIF requires a Source for each Factoid. In many projects, data that would comprise an IPIF Statement is given with no source (e.g. name, death, profession of a person"). In this case, the source is taken to be the project itself.
- TEI-XML personographies can express a factoid model by applying the @source and @resp attributes to the sub-elements of a <person> entry (see Schwarz 2020), but this is optional. Pragmatically, we suggest using the metadata of the TEI document as fallback (@source, @resp, ancestor::TEI/teiHeader/ titleStmt/editor etc.).
- To avoid ambiguity in combining statement filters in Person queries (does person/?place=Graz&role=professor mean "a professor in Graz" or "a professor, located in Graz for any reason"?), we defined a default behaviour (both conditions apply to the *same* Statement) and an optional parameter independentStatements=true to allow 'or' conditions. In this presentation, we will argue that such decisions are justified by the requirements of interoperability; and that our experiences in developing IPIF thus

far contribute usefully to debates surrounded data interoperability in the digital humanities.

Bibliography

Baillie, James. “Alternative Database Structures for Prosopographical Research”. *International Journal of Humanities and Arts Computing* 15, Nr. 1–2 (October 2021): 117– 32. https://doi.org/10.3366/ijhac.2021.0265.

Bradley, John Douglas. “A Prosopography as Linked Open Data: Some Implications from DPRR”. *Digital Humanities Quarterly* 014, Nr. 2 (29 July 2020). http:// digitalhumanities.org/ dhq/vol/14/2/000475/000475.html.

Brradley, John, and Harold Short. “Texts into Databases: The Evolving Field of New-Style Prosopography”. *Literary and Linguistic Computing* 20, Nr. Suppl (1 January 2005): 3–24. https://doi.org/10.1093/llc/fqi022.

Pasin, Michele, and John Bradley. “Factoid-Based Prosopography and Computer Ontologies: Towards an Integrated Approach”. *Literary and Linguistic Computing* 30, Nr. 1 (1 April 2015): 86–97. https://doi.org/10.1093/llc/fqt037.

Vogeler, Georg, Matthias Schlögl, and Gunter Vasold. “Data Exchange in Practice: Towards a Prosopographical API (Preprint)”. Paper given at BD2019 in co-location with RANLP 2019 in Varna (2019). https://doi.org/10.17613/YW4H-5F09.

Schwartz, Daniel L, Nathan P Gibson, and Katayoun Torabi. “Modeling a Born-Digital Factoid Prosopography Using the TEI and Linked Data”. *Journal of the Text Encoding Initiative*, 2020, 37.

Vasold, Gunter: Papilotte. A flexible and extensible IPIF server. https://github.com/ gvasold/papilotte (2019)

Zedlitz, Jasper. “Gedbas4all – New Data Model for Genealogy”. GenWiki, 2009. http:// wiki-en.genealogy.net/ Gedbas4all/Article = English version of Zedlitz, Jasper. „Gedbas4all – neues Datenmodell für die Genealogie“. *Computergenealogie*, Nr. 04 (2009).

Lexical Semantic Change in Literary Criticism

Haider, Thomas Nikolaus
thomas.haider@uni-goettingen.de
University of Göttingen, Germany; Max Planck Institute for Empirical Aesthetics, Frankfurt

Gittel, Benjamin
benjamin.gittel@uni-goettingen.de
University of Göttingen, Germany

Motivation

Approaches to the study of reading literature are increasingly attracting attention in the Digital Humanities (Lavin, 2020; Mavrody et al. 2021; Pianzola et al., 2020; Rebora et al., 2021). However, the historical change of literary practices has rarely been investigated on a quantitative basis, although social practices, and thus also literary practices, produce certain regularities of behavior, making them ideal for empirical study (Tuomela, 2002). Since the rules of literary practices are closely connected to literary concepts, we consider studying semantic change of concepts as a promising route to shed light on the change of literary practices over a long period of time.

In our contribution we investigate how the meaning of central concepts of literary criticism changed over the course of the long 19th century in the discourse of literary critics. Such concepts include ”author”, “character” (ger. Figur/Charakter), “literature”, “narrator” (Erzähler), “plot” (Fabel), “poetry” (Dichtung) or “story” (Geschichte). The central idea is that the semantic change of these concepts reflects fundamental change in literary practices, i.e., the practices of dealing with literary texts. The change in meaning of literary concepts particularly concerns (a) the reception of fictional texts and the so-called “fictive stance” (Lamarque and Olsen, 1994), (b) the categorization and interpretation of literary texts (Gittel, 2021), and (c) their evaluation (Friend, 2020).

Identifying a text as ”work of literature” or as ”poetry”, for example, evokes certain expectations and textual approaches on the part of literary critics, which are historically variable and in turn are reflected in changing word meanings (Rosenberg, 2017). Our working hypothesis is that by looking at longer periods of time, clusters of such word meanings are identifiable, which can then be interpreted qualitatively in terms of literary criticism’s history (Anz and Baasner, 2004; Habib, 2013; Hohendahl, 1985; Magerski, 2013) and further explored through quantitative analysis.

Approach

We employ methods from Natural Language Processing, specifically Lexical Semantic Change, to approach the diachronic change in meaning of literary concepts in an empirical manner. Our computational methodology is based on the task of diachronic word sense disambiguation and semantic change more generally (Schlechtweg et al., 2018;

Schlechtweg et al., 2019; Schlechtweg et al., 2020; Rother et al., 2020), encompassing both manual annotation and unsupervised methods from Distributional Semantics. We focus on diachronic contextualized embeddings (e.g., via BERT, see Laicher et al. (2021)), which we adapt to the historical literary domain.

As a basis for our analysis we collected a corpus of professional literary reviews from the popular German periodical "Die Grenzboten", which covers the years 1840–1921, a central period for the emergence of modern literary practices. In this corpus (30k texts), we manually annotated texts on whether they are reviews or not and implemented a heuristic to identify recurring journal sections which contain reviews, to build a high performing automatic classifier that identifies this type of text, which is crucial for a study of literary critics' practices. Furthermore, we acquire a second corpus of literary reviews that spans a wider time span (1750-1880), provided by the Austrian National Library.

Our analysis will focus on the change in meaning of the literary terms mentioned in the beginning. First of all, with a change point analysis, we aim to find changes in the similarity of concepts to themselves and each other over time, i.e., at which point a word becomes dissimilar to itself, or through which semantic fields selected concepts moved (Hamilton et al., 2016). Second, we aim to identify clusters of word meanings to disambiguate specific word senses and how the composition of clusters changed over time, i.e., whether certain senses emerge or vanish over time. To evaluate our (clustering) models, we carry out two main strategies, which are also suggested in the literature (Wevers and Koolen, 2020): (a) hypothesis testing, and (b) cluster coherence.

Regarding (a), we manually annotate categorical word senses (in context) for concepts like "Erzähler", for which we already know that it changed from meaning predominantly "author" to "fictive narrator". Regarding (b), we manually annotate via pairwise judgements, such that annotators are asked whether two instances of the same token in different contexts carry a similar or dissimilar meaning (whether they belong to the same cluster or not), and more generally, how similar two instances are on a continuous scale.

Bibliography

Thomas Anz and Rainer Baasner (ed.). 2004. Literaturkritik: Geschichte, Theorie, Praxis. München.

Stacie Friend. 2020. Categories of Literature. The Journal of Aesthetics and Art Criticism, 78(1):70–74.

Benjamin Gittel. 2021. Fiktion und Genre: Theorie und Geschichte referenzialisierender Lektürepraktiken 1870–1910. Berlin, Boston.

M. A. R. Habib (ed.). 2013. The Cambridge History of Literary Criticism, vol. 6 (The Nineteenth Century, c. 1830–1914). Cambridge.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Sstatistical Laws of Semantic Change. arXiv preprint arXiv:1605.09096.

Peter U. Hohendahl (ed.). 1985. Geschichte der deutschen Literaturkritik (1730-1980). Stuttgart.

Severin Laicher, Sinan Kurtyigit, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Explaining and Improving BERT Performance on Lexical Semantic Change Detection. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, 192–202, Online, April. Association for Computational Linguistics.

Peter Lamarque and Stein Haugom Olsen. 1994. Truth, Fiction, and Literature: A Philosophical Perspective.

Christine Magerski. 2013. Die Konstituierung des literarischen Feldes in Deutschland nach 1871: Berliner Moderne, Literaturkritik und die Anfänge der Literatursoziologie. Tübingen.

Matthew J. Lavin. 2020. Gender Dynamics and Critical Reception: A Study of Early 20th-Century Book Reviews from the New York Times. Journal of Cultural Analytics, 5(1).

Nika Mavrody, Laura B. McGrath, Nichole Nomura, and Alexander Sherman. 2021. Voice. Journal of Cultural Analytics 6 (2).

Federico Pianzola, Simone Rebora, and Gerhard Lauer. 2020. Wattpad as a Resource for Literary Studies. Quantitative and Qualitative Examples of the Importance of Digital Social Reading and Readers' Comments in the Margins. PloS one, 15(1): 1–46.

Simone Rebora, Peter Boot, Federico Pianzola, Brigitte Gasser, J Berenike Herrmann, Maria Kraxenberger, Moniek M Kuijpers, Gerhard Lauer, Piroska Lendvai, Thomas C Messerli, and Pasqualina Sorrentino. 2021. Digital humanities and Digital Social Reading. Digital Scholarship in the Humanities, 36 (Supplement 2): 230–250.

Rainer Rosenberg. 2017. Literarisch / Literatur. In Karlheinz Barck, Martin Fontius, Dieter Schlenstedt, Burkhart Steinwachs, and Friedrich Wolfzettel (eds.): Ästhetische Grundbegriffe, Vol. 3, 665–693. Stuttgart and Weimar.

David Rother, Thomas Haider, and Steffen Eger. 2020. CMCE at SemEval-2020 Task 1: Clustering on Manifolds of Contextualized Embeddings to Detect Historical Meaning Shifts. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, 187–193.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic Usage Relatedness

(DURel): A Framework for the Annotation of Lexical Semantic Change. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 169–174, New Orleans, Louisiana, June. Association for Computational Linguistics.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and Evaluating Lexical Semantic Change Across Times and Domains. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 732–746, Florence, Italy, July. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised Lexical Semantic Change Detection. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, 1–23, Barcelona (online), December. International Committee for Computational Linguistics.

Raimo Tuomela. 2002. The Philosophy of Social Practices: A Collective Acceptance View. Cambridge.

Melvin Wevers and Marijn Koolen. 2020. Digital Begriffsgeschichte: Tracing Semantic Change Using Word Embeddings. Historical Methods: A Journal of Quantitative and Interdisciplinary History, 53(4): 226–243.

"App-Solute News:" Comparison of Analog and Digital Mode in Newspaper Reading Between Intergenerational Teams

Hartinger, Teresa
teresa.hartinger@uni-graz.at
University of Graz, Austria

Marinšek, Urša
ursa.marinsek@uni-graz.at
University of Graz, Austria

Introduction

The project "App-Solute News" investigates intergenerational dynamics through reading newspapers and the transition of reading from analog to digital mode. Participants were tasked with creating digital stories reflecting on reading newspapers in intergenerational teams of two. In the stories, benefits and challenges of reading are explored. In addition, bringing forth images of age and the communication between younger and older adults the project is challenging stereotypes of aging.

Methodology

The main method in this project was digital storytelling, a pedagogical-narrative method whose outcomes are stories based on an exploration of a specific topic in a digital form. Participants create short narratives and combine them with various types of multimedia. Digital storytelling can be used in variety of settings with different agendas in mind. For example, it can be used to promote diversity, and reduce prejudices of aging through collaborations with creating images and narratives. It has already been applied in many projects in intercultural and educational contexts (e.g., "Mysty: Digital Storytelling" project).

The innovative aspect of the "App-Solute News" project is using digital storytelling in intergenerational collaborations between younger and older adults who created digital stories. The participants were students from the University of Graz, Austria who were paired with adults over the age of 60. Together they discussed reading newspapers in analog and digital forms. They worked with a print and an app version of *Kleine Zeitung*, the largest-circulation state daily newspaper, which is an example of a medium transitioning from analog to digital.

The teams exchanged experience about digitalization on an equal footing and obviated any notions of "expert" roles in the field of digitalization. Looking at how age plays a role in team interactions was important to the project, however, the participants were not made aware of the age aspect in detail in order not to influence the teamwork. The task was to develop digital stories and use them to create an image-based narrative that was recorded audio-visually. In their stories, the pairs discuss reading newspapers in the past and in the present, their perceptions, ideas, and positive and negative sides of analog and digital newspaper reading etc.

In addition, a questionnaire was used and employed after the digital stories had been created. The objective was to give the participants an opportunity to reflect on their overall experience. Apart from analyzing the participants' experience in the project, the main aim of the questionnaire was to see how intergenerationality and collaboration were perceived, and to see if the participants were challenged in terms of age(ing) stereotypes.

Findings

In this project, it appears that common age stereotypes, such as older people preferring the analog medium over the digital one and younger people preferring the digital medium, do not apply. Several people indicated that the choice of the preferred medium is not dependent on age, but on habits, interests, and individual needs.

It is evident that participants over 60 often prefer the app and become accustomed to it very quickly, or they had even used it before. Several participants under 35 years of age said they preferred the print version of the newspaper because they spend a lot of time in front of a screen at work. Although it was shown that the technical and digital tasks were taken on more by the group of people aged 35 and younger, the rest of the collaboration was determined by the diverse skills and personal preferences distributed differently in the age groups.

The analysis of the respondents' answers also showed that the intergenerational cooperation was experienced as consistently positive. Appreciation and collegiality as well as openness, interest and motivation played an equally important role as did creativity, harmony, and fun. Participants stated that they were able to learn from each other, they gained new social and digital competences, and developed their interpersonal skills. In many cases, they were surprised by the competence, needs and interests of their partners and stated that the cooperation was enriching.

The project goal of addressing age stereotypes and changing attitudes, behaviors and structures, as well as promoting new, different approaches and perspectives on age(ing), was achieved. Due to the inclusion of the aspect of intergenerationality, the analysis of the digital stories and the inherent societal ideas, fears, norms and values, regarding age and digitalization adds to the existing research of the connection between age and digitalization (e.g., Hausknecht et al., 2019; Loos et al., 2019; Weiß et al., 2017).

Bibliography

Hausknecht, S., Vanchu-Orosco, M., and Kaufman, D. (2019). Digitising the wisdom of our elders: Connectedness through digital storytelling. *Ageing and Society,* 39(12): 2714-2734.

Loos, E., Nimrod, G., and Fernández-Ardèvol, M. (Coords.). (2019). Older audiences in the digital media environment: A cross-national longitudinal study. *Wave 2 Report v1.0* . Montreal, Canada: ACT Project.

Mysty: Digital Storytelling. (2017).

Weiß, C., Stubbe, J., Naujoks, C., and Weide, S. (2017). *Digitalisierung für mehr Optionen und Teilhabe im Alter*. Bielefeld: Hans Kock Buch- und Offsetdruck GmbH.

Crowdsourcing as Collaborative Learning: A Participatory Annotation Project for the Photographic Materials of Shibusawa Eiichi

Hashimoto, Yuta
yhashimoto1984@gmail.com
National Museum of Japanese History, Japan

Kim, Boyoung
kim@shibusawa.or.jp
Shibusawa Eiichi Memorial Foundation

Nakamura, Satoru
nakamura@hi.u-tokyo.ac.jp
University of Tokyo

Kokaze, Naoki
xiao3feng10324@yahoo.co.jp
Chiba University

Inoue, Sayaka
s.inoue@shibusawa.or.jp
Shibusawa Eiichi Memorial Foundation

Shigehara, Toru
shigehara@shibusawa.or.jp
Shibusawa Eiichi Memorial Foundation

Nagasaki, Kiyonori
nagasaki@dhii.jp
International Institute for Digital Humanities

Introduction

Although crowdsourcing in the humanities has become common in the last decade (Hedges and Dunn, 2017; Terras, 2013) it is still challenging for most humanities scholars and academic institutions to conduct research successfully

because they must draw public attention to their project, keep the participants engaged and motivated, and pay close attention to the quality of the outcome.

Nevertheless, the subjects of humanities crowdsourcing can also serve as educational resources for humanities studies. We hypothesize that this characteristic can be used to resolve crowdsourcing difficulties. In this study, we conducted a crowdsourcing project to test this hypothesis, which aims to annotate historical photographs related to Shibusawa Eiichi.

Background and Aims of the Project

Shibusawa Eiichi [1] (1840-1931, Fig. 1) was a leading figure in the development of modern Japan, who introduced many financial and economic reforms as a member of the Ministry of Finance in the fledgling Meiji government. After leaving the government in 1873, he ventured into the business world and established a wide variety of companies and economic organizations, including the First National Bank, the Tokyo Chamber of Commerce, and the Tokyo Stock Exchange. He was involved in the foundation of roughly 500 companies and 600 non-profit organizations in the fields of education, social welfare, and health.

The Shibusawa Eiichi Memorial Foundation compiled primary sources related to Shibusawa's life and achievements, including his diaries, letters, and newspaper articles, and published them in 68 volumes as *Shibusawa Eiichi Denki Shiryo* (Shibusawa Eiichi Biographical Materials) [2]. Among these volumes, supplementary volume 10 was devoted to photographic materials and contained over 700 photos of people, monuments, documents, buildings, and landscapes in which Shibusawa was involved (see Fig. 2).

Our project is part of the ongoing efforts of the Foundation to digitize *Shibusawa Eiichi Denki Shiryo*, which aims to enrich the metadata of the digitized photos in supplementary volume 10 through crowdsourcing. As Shibusawa had been involved with many important businesspersons and was related to a wide variety of economic organizations and legislation, digitizing these photographs in a structured way will make a considerable contribution to the field of modern Japanese history.



Fig. 1
A portrait of Shibusawa Eiichi (1840-1931)


164 栄一と三井組の人々

第一国立銀行は明治六年六月十一日創立総会を開いた。三井・小野・島田を主軸として一般に株式を募集したものである。上掲写真、前列右より。三野村利左衛門。栄一。三井高福。斎藤純造。永田甚七。後列右より。三井高喜。三井高朗。第一国立銀行本店は三井組為替座として海運橋畔に建築された建物を買取って開業した。写真は建築中の為替座である。新聞に広告して広く株式を募集した。

165 建築中の第一国立銀行

166 銀行発祥の地碑
（渋沢敬三筆。昭和38年６月建立）

167 株式募集広告（東京日々新聞。明治６年１月７日）

53

Fig. 2
Photogpraphs in Supplementary Volume 10.

Method

To resolve the difficulties of academic crowdsourcing, we designed our system based on the concept of “collaborative learning” in pedagogy. Collaborative learning refers to a situation or environment in which multiple people learn or attempt to learn something together (Dillenbourg, 1999). Unlike individual learning, in collaborative learning, people capitalize on one another's resources and skills.

Our fundamental idea is to brand our crowdsourcing system as a place for collaborative learning in which participants share their knowledge related to the historical photographs by performing annotation tasks. A similar attempt has already been made to the transcription of historical Japanese documents with considerable success (Hashimoto et al., 2018). However, it is unknown whether the same approach can be applied to photographic materials or not.

Implementation

Our crowdsourcing system, *Minna de Koshashin* (let’s annotate historical photographs together; https://denkiphoto.shibusawa.or.jp/annotate/), is a single-page application (SPA) built with Vue.js [3] and Firebase [4]. The participants of *Minna de Koshashin* are invited to browse the photos of Shibusawa Eiichi and to conduct the following six tasks on each photo:

1. Identifying a person
2. Identifying a location
3. Transcription of textual information
4. Uploading a current photo of a place
5. Registration of bibliographic information
6. Writing a comment.

Since Tasks 4-6 are open-ended tasks that cannot be determined as completed, we use Tasks 1-3 to track the progress of our project; The annotation of a photo is considered completed when all of Tasks 1-3 are completed. Deciding which task to perform is up to each participant.

To ensure that collaborative learning is achieved through the above tasks, we implemented the following four features:

1. **Task history**. When conducting a task, participants are required to write the reason for their decision, which will be recorded in the “history” tab (see Fig. 3) and can be referenced by others. This process makes the tacit knowledge of participants explicit, thus facilitating mutual learning among participants.
2. **Activity sharing**. All the activities of participants are shared on the “timeline” view in a chronological order (see Fig. 4), promoting horizontal propagation of knowledge and providing social stimuli.
3. **Communication and feedback**. Participants can seek advice and help from others in the “discussion” tab (see Fig. 5) or use the comment function. Anyone can correct the annotations of the other participants. Through communication and collaboration among participants, annotations undergo multiple checks, leading to improved output quality.
4. **Mutual evaluation**. The system has a leaderboard to stimulate participants' competitive spirit, but the rank is determined not by the number of tasks but by the number of "likes" received from others, inducing altruistic behavior.



Fig. 3
History tab

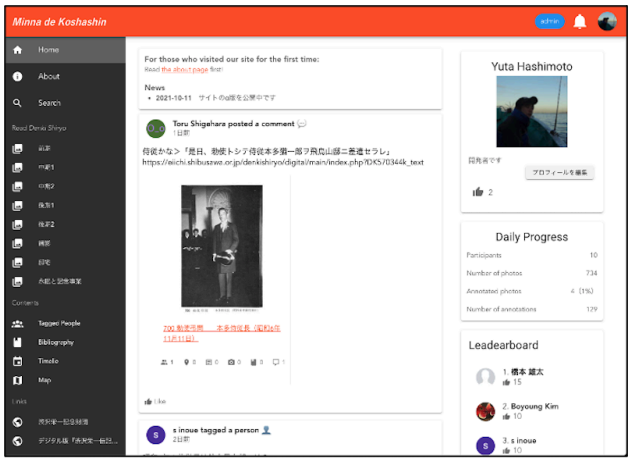


Fig. 4
Timeline view



Fig. 5
Discussion tab

Results

Minna de Koshashin, as an outcome of our project, was launched in December 2021. As of April 2022, four months after the system launch, 59 people have created user accounts on our website. A total of 1,405 annotations have been created. The breakdown is described in Table 1. 159 photos out of 734 (22%) are marked as "completed". Although we have not been able to recruit a large number of participants, a small number of dedicated participants have continually involved in performing annotations.

Type of annotation	Count
Person	937
Location	242
Transcription	26
Photo upload	26
Bibliographic Information	131
Comment	43
Total	1405

Table 1
Breakdown of the annotations according to types

Conclusion

Our project has just entered the operational phase. Still, our approach based on collaborative learning seems to be working to some extent. The majority of participants provide bibliographic information as the basis for their annotations, allowing others to learn about the information. On the other hand, communication and discussion among participants is not very active. This may change as the user base expands.

In any case, further research is needed to evaluate the effectiveness of our approach. We plan to conduct further investigations, such as assessment of annotation qualities and interviews with participants.

Bibliography

Hedges, M. and Dunn, S. (2017). *Academic crowdsourcing in the humanities: Crowds, communities and co-production*. Chandos Publishing.

Terras, M. (2013). Crowdsourcing in the digital humanities. In book: *A New Companion to Digital Humanities*, pp. 420-439.

Dillenbourg, P. (1999). What do you mean by collaborative learning? In book: Collaborative-learning: Cognitive and Computational Approaches, pp. 1-19.

Hashimoto, Y., et al. (2018). Minna de Honkoku: Learning-driven Crowdsourced Transcription of Pre-modern Japanese Earthquake Records. In *Book of Abstracts, Digital Humanities 2018*, Mexico City, pp. 207-210, https://dh2018.adho.org/en/minna-de-honkoku-learning-driven-crowdsourced-transcription-of-%E2%80%A8pre-modern-japanese-earthquake-records/ (accessed April 18, 2022).

Notes

1. Shibusawa is the surname.
2. Available online at https://www.shibusawa.or.jp/english/eiichi/denki_shiryo.html.
3. Vue.js. https://vuejs.org/.
4. Firebase. https://firebase.com/.

An extensible Cor framework for textual publication and structure.

Hayward, Nicholas John
nhayward@luc.edu
Loyola University Chicago, United States of America

An extensible *Cor*framework provides a structure and core for the development, analysis, and textual publication of polyglot research and projects. Originally developed as the underlying structure of the Woolf Online (Caughie et al., 2013)project's *Mojulem*framework, it has continued to evolve to support multiple project instances[1], each internally cross-referenced and networked to enable data and software analysis.

As a publication tool, the *Cor*framework builds on the concept of 'knowledge sites', suggested by Peter Shillingsburg (Shillingsburg, 2006), supplementing a core publication framework with modules such as OCR, editors, image viewers, data analysis, and software metrics.

Mojulem, for example, initially required four underlying core structures, which included *CorCode*, *CorForm*, *CorPix*, and *CorTex.*
This development has continued with the Verne Digital Corpus (Hayward, 2017)and the NSF funded Metrics Dashboardproject (Shilpika et al., 2015) to include the addition of *CorAssess*and associated software metrics to the underlying *Cor*framework.

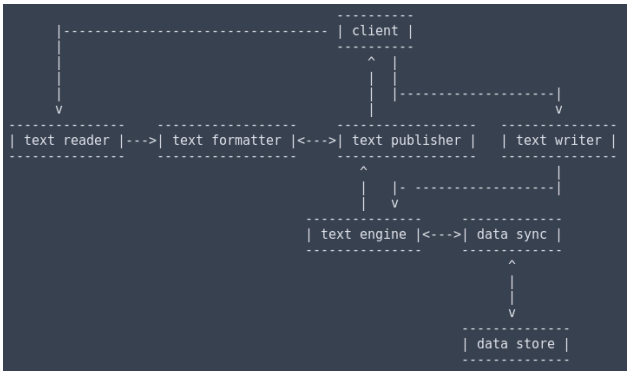
CorTex

The core structure of this framework is a *CorTex*, a stable resource containing merged or compacted plain text transcriptions of a work's variant expressions. It stores all information about text and variations, ready to be extracted for display of variation amongst versions; it is not necessary to recompute them. *CorTex*is the entity all standoff properties (markup, annotations, links...) points, providing system stability, each version's text, and variations from other texts. Stability and endurance of *CorTex*is protected by multiplying duplicate copies locked with a digital signature, which verifies for each user that a *CorTex*copy is viable. Analysis of *CorTex* variable forms provides statistical feedback to guide production of a conflated text, for example with the English language translations of a given author's edition. Whilst these statistical results are no guarantee of an ultimately correct translation, in the example of multiple language editions, they do offer a conflated text with the highest viable agreement amongst collated texts. Textual disagreements are currently resolved by assigning probability values, a higher value defining a greater probability of accuracy and agreement amongst the collated texts. Probability results provide an initial filter of problematic passages in each translation to conflate a text with the highest probability of agreement amongst the translations per edition.

These results can then be provided for further research and assessment, and act as a suitable starting guide for further analysis of the conflated text and, where applicable, translation.

The CorTex provides an abstracted core for text management, processing, and modularised structure within a variety of aggregated projects.

An initial structure for a project framework, for example, may be considered as follows,



Such pathways correspond to common data patterns defined and observed for frameworks based upon a general

CorTex. For example, we may clearly identify and define a separation of concerns for a user's request for data, perhaps a chosen page for the current selected edition, from the act of reading the text, formatting it for rendering to the user, and the final act of publication. Each internal pipeline has a clear focus of purpose, thereby removing tightly coupled components, and enabling secure use of the underlying textual data.

Text Engine

The abstracted *text engine*, for example, provides a kernel for supporting I/O (input/output) requests, secure access to write to datastore persistency, whilst enabling subsequent publication of queried or cached sync data.

The *text engine*supports plain text by default, abstracting support for multiple variant formats using parsers for text I/O. Such parsers may be integrated at a higher level in a chosen pipeline, thereby simplifying the process role of the *text engine*. All revisions to texts may be signed as updates by the text engine, enabling editorial actions to be easily recorded in a time-series, sequential manner. The *text engine*, therefore, creates a clear separation of concerns for textual markup styles, version history, authorial record, and associated metadata. Textual records maintain clear integrity from draft to proofs to publication, including separate options to redraft or revise a given data record, ensuring integrity for both the original draft and proofs copy of the textual data.

A notable benefit of this approach, in addition to clear data integrity and modularity of design, is the option to define an API (application programming interface) for each constituent component, from the *text engine*to the *text reader*. This abstraction of structure and design enables variant client options as well, from dynamic implementations of client rendered content to static, single query text publication. Publication channels are incorporated to enable a variety of secure queries from the *client*to the *text engine*.

Summary

This paper will introduce the structure and development of the underlying *Cor*framework with specific focus on a *CorTex*, its development, and associated *text engine*use within a publication framework.

Bibliography

Caughie, Pamela L., Hayward, Nicholas J., Hussey, Mark., Shillingsburg, Peter L., and Thiruvathukal, George. K., eds. (2013). *Woolf Online*. Web. http://www.woolfonline.com.

Goodman, N. (1976). *Languages of Art*. Hackett Publishing.

Hayward, Nicholas. J. (2017). *A Cor infrastructure for textual analysis - From Woolf to Verne*. DH2017 Conference, McGill University and Université de Montréal, Montreal, Canada

Shillingsburg, P. L. (2006). *From Gutenberg to Google: Electronic Representations of Literary Texts*. Cambridge University Press.

Shilpika; Thiruvathukal, George K.; Aguiar, Saulo; Läufer, Konstantin; and Hayward, Nicholas J. (2015). *Software Metrics and Dashboard*, https://ecommons.luc.edu/cs_facpubs/87/

Thiruvathukal, George K., Hayward, Nicholas. J., Laüfer, Konstantin., and Shilpika. (2018). *Metrics Dashboard: A Hosted Platform for Software Quality Metrics*, arXiv:1804.02053 [cs.SE].

Thiruvathukal, George K., Hayward, Nicholas. J., Laüfer, Konstantin. Shilpika, and Aguiar, Saulo. (2015). *Towards Sustainable Digital Humanities Software*, Chicago Colloquium on Digital Humanities and Computer Science.

Notes

1. Current projects include Malory Project (www.maloryproject.com), Modernist Magazines (www.modernistmagazines.com), and Woolf Online (www.woolfonline.com).

Why arts and humanities publications get retracted: a topic modeling analysis of the retraction notices

Heibi, Ivan

ivan.heibi2@unibo.it

Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy; Digital Humanities Advanced Research Centre (/DH.arc), Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy; Research Centre for Open Scholarly Metadata, Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

Peroni, Silvio
silvio.peroni@unibo.it
Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy; Digital Humanities Advanced Research Centre (/DH.arc), Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy; Research Centre for Open Scholarly Metadata, Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

Background

The retraction of a scholarly peer-reviewed publication means that its corresponding venue (e.g., journal) has decided to withdraw it due to some irregularities/errors. A retraction could be partial or full. In case of a partial retraction, articles have flawed data or content errors in small parts, and the correction of the erroneous article portions keeps the general information and the article's stated conclusions uncompromised. On the other hand, a full retraction is *"a mechanism for correcting the literature and alerting readers to articles that contain such seriously flawed or erroneous content or data that their findings and conclusions cannot be relied upon"* (Barbour et al., 2009). Partial retractions are not helpful and cannot determine the status of a publication, therefore it is more reasonable to focus on definitive full retractions.

Most retractions occur in STEM (Science, technology, engineering, and mathematics), while social sciences and humanities relatively account for a small portion compared to these fields (Vuong et al., 2020). Reasons for retraction are mainly classified in two categories: (a) honest error and (b) misconduct. When a retraction is raised, the venue needs to formally accompany such decision with retraction notice – a document that provides sufficient information about the reason for retraction and why the findings are considered unreliable, in addition, should explicitly mention whether this was due to misconduct or an honest error. Retraction notices should be freely available and linked to the retracted article in both the PDF and online version (Barbour et al., 2009).

Several studies worked on studying the reasons for retraction. High attention has been given to STEM and mainly to health science (Li et al., 2018). Few past studies worked on the analysis of the arts and humanities domain (a rare example is the work of Halevi (2020) and our work (Heibi & Peroni, 2021).

Aim

Considering the less attention that has been given in the study of the retraction phenomenon and in particular to the reasons of retraction in the arts and humanities domain, the aim of this work is to investigate the reasons of retraction in the arts and humanities through the automatic analysis of the retraction notices and a comparison of these results with the data provided by other services which have worked on labeling such reasons.

Method

Our methodology is articulated in three main steps:

1. Gathering the retraction notices of all the retracted articles in arts and humanities using the Retraction Watch Database (http://retractiondatabase.org/) or by querying large bibliographic databases (e.g., ScienceDirect by Elsevier) searching for terms such as "RETRACTED" and subsequently getting the retraction notice.
2. Building and running a topic modeling (TM) analysis using the Latent Dirichlet Allocation (LDA) model (Jelodar et al., 2019) giving by input a corpus containing the full text of all the retraction notices collected in step 1. This process is done using MITAO, a user-friendly interface to create a customizable visual workflow for text analysis (Ferri et al., 2020; Heibi et al., 2021).
3. Analyzing the topic modeling outcomes using dynamic visualizations which help us observe and investigate the results as a function of other features and compare our findings with the data/annotations provided by Retraction Watch (Marcus & Oransky, 2014), a manual-curated database collecting retractions from several domains.

Expected results

The study we have presented in this abstract is a work in progress analysis. The ambition is to uncover and infer new insights regarding the reasons of retraction in the arts and humanities domain which did not emerge in the past studies – e.g., the work of Halevi (2020) – or in the annotations of services such as Retraction Watch. We can hypothesis that is very plausible to find out that the outcomes of our analysis will be related to retraction reasons such as "plagiarism" and "content duplication", which are the most recurring reasons following the annotations of Retraction Watch (also demonstrated by the work of Halevi (2020)). The fact that these reasons are the most popular ones for the

humanities domain, is due to the fact that the detection of these forms of retraction are well defined and less prone to interpretation (Dhammi & Ul Haq, 2016), therefore easier to establish considering the different interpretable facets of the humanities arguments.

In addition, an aspect to investigate concerns the evaluation of the reliability of a TM analysis in the classification of misconduct and honest error reasons compared to the human annotations provided by Retraction Watch. This study might also evaluate the TM technique through the consideration of other text analysis methodologies, e.g., SBERT (Reimers & Gurevych, 2019).

Bibliography

Barbour, V., Kleinert, S., Wager, E. and Yentis, S. (2009). *Guidelines for Retracting Articles*. Committee on Publication Ethics doi: 10.24318/cope.2019.1.4. https://publicationethics.org/node/19896 (accessed 13 November 2020).

Dhammi, I. and Ul Haq, R. (2016). What is plagiarism and how to avoid it?. *Indian Journal of Orthopaedics*, **50**(6): 581 doi: 10.4103/0019-5413.193485.

Ferri, P., Heibi, I., Pareschi, L. and Peroni, S. (2020). MITAO: A User Friendly and Modular Software for Topic Modelling. *PuntOorg International Journal*, **5**(2): 135–49 doi: 10.19245/25.05.pij.5.2.3.

Halevi, G. (2020). Why Articles in Arts and Humanities Are Being Retracted?. *Publishing Research Quarterly*, **36**(1): 55–62 doi: 10.1007/s12109-019-09699-9.

Heibi, I., Peroni, S., Pareschi, L. and Ferri, P. (2021). MITAO: a tool for enabling scholars in the humanities to use Topic Modeling in their studies. Text Alma Mater Studiorum - Università di Bologna doi: 10.6092/UNIBO/AMSACTA/6712. http://amsacta.unibo.it/6712 (accessed 19 July 2021).

Heibi, I. and Peroni, S. (2021). A quantitative and qualitative citation analysis of retracted articles in the humanities. *ArXiv:2111.05223 [Cs]* http://arxiv.org/abs/2111.05223 (accessed 10 November 2021).

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y. and Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, **78**(11): 15169–211 doi: 10.1007/s11042-018-6894-4.

Li, G., Kamel, M., Jin, Y., Xu, M., Mbuagbaw, L., Samaan, Z., Levine, M. and Thabane, L. (2018). Exploring the characteristics, global distribution and reasons for retraction of published articles involving human research participants: a literature survey. *Journal of Multidisciplinary Healthcare*, **Volume 11**: 39–47 doi: 10.2147/JMDH.S151745.

Marcus, A. and Oransky, I. (2014). What Studies of Retractions Tell Us. *Journal of Microbiology & Biology Education*, **15**(2): 151–54 doi: 10.1128/jmbe.v15i2.855.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *ArXiv:1908.10084 [Cs]* http://arxiv.org/abs/1908.10084 (accessed 15 April 2022).

Vuong, Q.-H., La, V.-P., Ho, M.-T., Vuong, T.-T. and Ho, M.-T. (2020). Characteristics of retracted articles based on retraction data from online sources through February 2019. *Science Editing*, **7**(1): 34–44 doi: 10.6087/kcse.187.

Patterns of Verb Usage in Immanuel Kant's Critical Writings

Heßbrüggen-Walter, Stefan
shessbru@hse.ru
HSE University, Moscow, Russian Federation

Fischer, Frank
fr.fischer@fu-berlin.de
Freie Universität Berlin, Berlin, Germany

Meier-Vieracker, Simon
simon.meier-vieracker@tu-dresden.de
Technische Universität Dresden, Dresden, Germany

Verbs have not been a prominent object of investigation in the digital humanities, although we do find computational literary studies on acoustic phenomena in novels (Katsma 2014) or narrativity of stage instructions in modernist drama (Trilcke et al. 2020) that both focus on the use of verbs in the respective corpora. Philosophical discussions about verbs often refer only to certain classes that are considered philosophically relevant in a given context: illocutionary verbs (Green 2021), verbs expressing propositional attitudes (Nelson 2019), or intensional transitive verbs (Forbes 2020). Our approach is both more comprehensive and more specific. We look at finite verbs regardless of their semantic classification (in this we follow Langer 1927). At the same time, our interest is 'philological' in that we aim to understand how verbs contribute to the meaning and interpretation of historical philosophical texts (Kahn 2003 follows a similar approach, however limited to only one verb, 'to be'). Whereas most philosophical enquiry focuses on nouns as the linguistic side of relevant concepts like reason, duty, taste, etc., the study of verbs that function as predicates and thus relate these concepts can complete the

picture of how philosophical judgements and arguments are made.

We present first results of an investigation of verb usage in Kant's major critical writings. This corpus has the advantage that these texts present a unified system, so that we can ignore their diachronic dimension. And yet they allow for a contrastive analysis, because the three philosophical subdisciplines theoretical philosophy, practical philosophy and aesthetics are clearly mirrored in the structure of this corpus. We exclude minor writings published after the first edition of the *Critique of Pure Reason* (1781), since they either cannot be clearly assigned to one of the three subdisciplines of philosophy under investigation or belong to subdisciplines such as philosophy of history, political philosophy or philosophy of religionwhich form only a small part of Kant's overall critical system. The goal of our analysis consists in the identification of verbs that are typical for the respective subcorpus and philosophical subdiscipline. We aim to show that Kant's usage of verbs differs depending on the philosophical subdiscipline the respective text belongs to.

The corpus consists of the main writings of Kant's critical philosophy and is divided into three subcorpora (tab. 1): 1) *Prolegomena*, *Metaphysical Foundations of Natural Science*, and the second edition of the *Critique of Pure Reason* (the inclusion of both editions would have introduced a lack of balance in the dataset) in theoretical philosophy, 2) *Foundations of the Metaphysics of Morals*, *Critique of Practical Reason* and *Metaphysics of Morals* for practical philosophy and 3) the *Critique of Judgment* for aesthetics and teleology.

Assembled Subcorpora	Works	N tokens
Theoretical Phil. (3 works)	*Prolegomena*	52,588
	Metaphysical Foundations of Natural Science	41,962
	Critique of Pure Reason	216,587
	Total	**311,137**
Practical Phil. (3 works)	*Foundations of the Metaphysics of Morals*	32,958
	Critique of Practical Reason	67,090
	Metaphysics of Morals	106,423
	Total	**206,471**
Aesth. / Tel. (1 work)	*Critique of Judgment*	**127,939**

Tab. 1 number of tokens per work and subcorpus

The digital edition we used employs modernised orthography which increases the reliability of verb identification through POS-tagging. The texts were tagged and lemmatized with the Stanza POS tagger (Qi et al. 2020) with some manual post processing. We then calculated the key verbs (Culpeper/Demmen 2015) for each subcorpus in contrast to the complete critical writings as reference corpus. We used log-likelihood ratio (Dunning 1993) as keyness measure which can handle the differences in the subcorpus sizes and results in a list of verbs that are used significantly more often in a subcorpus than would be expected from a hypothetical equal distribution

Results

With the help of our domain knowledge, we can state that Kant's verb usage shows clear differences across the three subdisciplines. Moreover, we can identify areas within the respective subdiscipline in which verbs make a substantial semantic contribution to Kant's philosophical language.

In practical philosophy, many of the high-ranking verbs belong to the semantic field of law (including the moral law, i. e. the Categorical Imperative): to acquire (*erwerben*), to obligate (*verpflichten*), to force (*zwingen*). Others are generic terms for actions (*handeln, machen, tun*). Only one verb denotes an emotion (to love, *lieben*). In theoretical philosophy, the two most high-ranking verbs are associated with the faculty of sensibility: to give, *geben*, associated with what is given in sensibility, and to intuit (*anschauen*). Others seem to belong to natural philosophy (*erfüllen*, to fill, e. g. space), to move (*bewegen*), to begin (*anfangen*), to change (*verändern*). Some are what we could call 'generic ontological verbs', to take place (*stattfinden*), to exist *(existieren*). Only one verb is connected to a pertinent epistemic activity, to construe *(konstruieren*). In aesthetics and teleology, i. e. in *Critique of Judgment*, verbs that express an activity are more prominent: judging (*urteilen*, *beurteilen*) plays, of course, an eminent role as do verbs that denote an aesthetic response (to please *, gefallen*, to entertain (*unterhalten*), the communicative force of an aesthetic judgment (to require *, ansinnen*) or the act of communication itself (to communicate *, mitteilen*).

Further research will be required to investigate the syntactic diversity of Kant's use of verbs (finite verb forms compared to participles or infinitives) and its relation to 18th century German in general. Moreover, the collection of typical verb-noun collocations as *Recht erwerben* (to aquire a right) will be a useful step.

Our corpus, code and data will be published under free licenses.

Bibliography

Culpeper, Jonathan / Demmen, Jane (2015). "Keywords." In: Biber, Douglas / Reppen, Randi (eds.): The Cambridge Handbook of English Corpus Linguistics. Cambridge: Cambridge University Press. 90–105. (doi:10.1017/ CBO9781139764377.006)

Dunning, Ted (1993). "Accurate methods for the statistics of surprise and coincidence." In: Computational Linguistics 19 (1), S. 61–74.

Forbes, Graeme (2020), "Intensional Transitive Verbs", *The Stanford Encyclopedia of Philosophy* (https://plato.stanford.edu/archives/win2020/entries/intensional-trans-verbs/).

Green, Mitchell (2021), "Speech Acts", in: *The Stanford Encyclopedia of Philosophy*, (https://plato.stanford.edu/archives/fall2021/entries/speech-acts/).

Kahn, Charles H. (2003) The verb "be" in ancient Greek, Indianapolis: Hackett.

Katsma, Holst (2014). "Loudness in the Novel." (= *Stanford Literary Lab, Pamphlet* Nr. 7 [September 2014].)

Langer, Susanne K. (1927). "A Logical Study of Verbs." In: *The Journal of Philosophy.* Band 24, Heft 5 (März 1927), S. 120–129. (doi:10.2307/2015082)

Nelson, Michael (2019) "Propositional Attitude Reports", *The Stanford Encyclopedia of Philosophy*, (https://plato.stanford.edu/archives/spr2019/entries/prop-attitude-reports).

Qi, Peng / Zhang, Yuhao / Zhang, Yuhui / Bolton, Jason / Manning, Christopher D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages.In Association for Computational Linguistics (ACL) System Demonstrations. 2020.

Trilcke, Peer / Kittel, Christopher / Reiter, Nils / Maximova, Daria / Fischer, Frank (2020). "Opening the Stage: A Quantitative Look at Stage Directions in German Drama." In: *DH2020: »carrefours/intersections«.* 22–24. Juli 2020. Conference Abstracts, University of Ottawa.

Worlding databases: A decolonising approach to the structuring and representation of data about global arts

Hidalgo Urbaneja, Maribel
m.hidalgourbaneja@arts.ac.uk
University of the Arts London, United Kingdom

Velios, Athanasios
a.velios@arts.ac.uk
University of the Arts London, United Kingdom

Goodwin, Paul
paul.goodwin@arts.ac.uk
University of the Arts London, United Kingdom

Worlding Public Cultures: The Arts and Social Innovation is an international multi-partner project that proposes *worlding* (Heidegger, 2002 [1950]; Spivak, 1985; Hunt, 2014; Cheah, 2016) as an activating concept and analytical tool. The concept goes beyond current top-down models of "inclusion," "diversity" and other representations of the "global". *Worlding* grounds the global within local worlds and allows entangled histories to emerge, opening pathways to decolonise "universal" Western narratives and epistemologies. Practices that generate new forms of knowledge in the digital sphere or challenge existing ones are seen as a *worlding* exercise by decolonial digital humanities (Risam, 2018).

One of the main outputs of the *Worlding Public Cultures* project is a dedicated website, https://www.worldingcultures.org/, that will constitute a hub for exchange of information about decolonising activities led by museums, universities, and other cultural and activist organisations. A key element on that website is the publicly accessible database that will share a structured set of curated data about how "global" narratives are being told and shared by exhibitions, academic courses, public events, and activist initiatives around the world. The database can be queried by place, time span, actors, and topics. A team of researchers from different backgrounds based in multiple geographical locations (London, UK; Amsterdam, Netherlands; Heidelberg, Germany; Montreal and Ottawa, Canada) collects data about localised events and activities. The database will provide museum professionals, scholars, teachers and students, and cultural activists with information that can help them when planning and organising activities or projects about global arts and culture.

Gathering and curating data about the "global" increases the visibility of non-canonical and non-western arts and cultures but *worlding* requires an effort that goes beyond representation. Developing a database in the context of the Worlding Public Cultures project implies the use of *worlding* as a concept and tool to rethink and critique the epistemological foundations of databases, ontologies, and structured vocabularies. The database structure is mapped on the CIDOC Conceptual Reference Model and database entries often include terminology from structured vocabularies and authority files such as the Getty Vocabularies and the Virtual International Authority File (VIAF). The adoption of well-established and "universally" recognised ontologies and vocabularies is considered good

practice for data integration and exchange, yet, it reinforces existing power dynamics and knowledge biases. Moreover, the collected data that will populate the database has been modelled on existing standards and canons.

This paper will shed light on the research questions that guide the ongoing process of developing the Worlding Public Cultures database and collection of the data that will populate it. Key areas of discussion encompass: the theoretical interrogation of the strategies, and actions frameworks that have given shape to databases, ontologies, and structured vocabularies; the engagement with professional communities responsible for the formulation of the data ontology used in the project; the ethical and critical aspects taken into consideration when reproducing data collected from multiple sources; and the design of a graphical interface within typical web applications that exposes the cultural and epistemological biases implicit in the CIDOC CRM, Getty Vocabularies, and VIAF to users that interact with the database.

Bibliography

Bruseker, G. and Guillem A. (2018). Decolonialism and Formal Ontology: Self-critical Conceptual Modelling Practice. *DH2018 Conference Proceedings*. https://dh2018.adho.org/decolonialism-and-formal-ontology-self-critical-conceptual-modelling-practice/ (accessed 20 April 2022)

Duarte, M. E. and Belarde-Lewis, M. (2015). Imagining: Creating Spaces for Indigenous Ontologies, *Cataloging & Classification Quarterly*, 53(5–6), pp. 677–702. doi: 10.1080/01639374.2015.1018396.

Heidegger, M. (2002). The Origin of the Work of Art. In *Martin Heidegger: Off the Beaten Track,* edited and translated by J. Young and K. Haynes, 1–56. Cambridge: Cambridge UP.

Hunt, S. (2014). Ontologies of Indigeneity: The Politics of Embodying a Concept. *Cultural Geographies,* 21(1): 27–32. DOI: 10.1177/1474474013500226.

Krmpotich, C. and Somerville, A. (2016). Affective Presence: The Metonymical Catalogue. *Mus Anthropol*, 39: 178-191. https://doi.org/10.1111/muan.12123 (accessed 20 April 2022)

Risam, R. (2019). New digital worlds: Postcolonial digital humanities in theory, praxis, and pedagogy. Evanston, Illinois: Northwestern University Press.

Shep, S., Frean, M, Owen, R., Pope, R. Reihana, P., Chan, V. (2020). Indigenous frameworks for data-intensive humanities: recalibrating the past through knowledge engineering and generative modelling. https://hal.inria.fr/hal-02461884v2 (accessed 20 April 2022)

Spivak, G. C. (1985). Three Women's Texts and a Critique of Imperialism. Critical Inquiry 12(1) "Race," Writing, and Difference (Autumn): 243–261.

Srinivasan, R. (2013). 'Re-thinking the cultural codes of new media: The question concerning ontology', *New Media & Society*, 15(2), pp. 203–223. doi: 10.1177/1461444812450686.

Turner, H. (2020). *Cataloguing Culture. Legacies of Colonialism in Museum Documentation*. UBC Press.

Modelling the relationship between morphosyntactic features and discourse relations in a multimodal corpus of primary school science diagrams

Hiippala, Tuomo
tuomo.hiippala@helsinki.fi
University of Helsinki, Finland

Haverinen, Jonas
jonas.haverinen@helsinki.fi
University of Helsinki, Finland

In 1998, Watanabe and Nagao published a pioneering article on the relationship between written language and pictorial representations in diagrams (Watanabe and Nagao 1998). By manually analysing 31 diagrams from Japanese books of flora that describe the shape, features and environment of plants, Watanabe and Nagao showed that morphosyntactic features of textual elements could be mapped to specific discourse relations that held between the text and pictorial representation of a plant. They also formulated a set of rules to support the computational processing of diagrammatic representations, which could be used to infer what kinds of relations hold between textual and pictorial elements.

From a contemporary standpoint, the diagrams studied by Watanabe and Nagao (1998) can be approached from the perspective of multimodality theory, which studies how human communication relies on intentional combinations of multiple "modes" of expression (Bateman et al. 2017). From a multimodal perspective, individual diagrams may be treated as instances of the diagrammatic semiotic mode, which integrates natural language and diverse visual expressive resources into a common discourse organisation (Hiippala and Bateman 2021). Against this backdrop, the rules formulated by Watanabe and Nagao (1998) can be treated as descriptions of their multimodal discourse

structure, which guide the viewers towards interpretations of what combinations of modes mean in their context of occurrence (Bateman 2020).

In this contribution, we revisit the work of Watanabe and Nagao (1998) using a recently published multimodal corpus of 1000 primary school science diagrams in English. This openly-available corpus, named AI2D-RST, contains multiple layers of cross-referenced annotations for expressive resources, layout and discourse structure, which have been created by trained experts (Hiippala et al. 2021). Our aim is to establish whether a similar mapping between morphosyntactic features and discourse relations proposed by Watanabe and Nagao (1998) can be found in English-language diagrams that serve similar communicative goals, that is, depict and explain various natural phenomena. Acknowledging the multimodal nature of diagrams, we also complement the morphosyntactic features with information about diagram layout and use of lines.

In contrast to the manual analysis in Watanabe and Nagao (1998), we adopt a corpus-driven approach to examine discourse relations between textual and pictorial elements. We extract 2580 discourse relations from the AI2D-RST corpus that hold between pictorial and textual elements, focusing on relations that name entire objects ("identification") or describe part-whole relations ("elaboration"). We extract the following features for each pair of elements: (1) whether the textual element consists of a nominal, clause, modifier or numeral, (2) the distance between elements in the layout, (3) the angle between pictorial and textual elements, and (4) whether the elements are connected using a line.

We use the aforementioned features to train a random forest classifier with 10 decision trees to predict whether the textual element names or describes a pictorial element. We use 10-fold cross-validation to evaluate the classifier, which achieves an average macro F1-score of 0.86 (standard deviation: 0.06). An analysis of how much each feature contributes to classification decisions reveals that apart from numerals, linguistic information is largely irrelevant. The distance between the pictorial and textual elements and whether they are connected using a line are the most important features for determining the function of a text element (see Figure below).

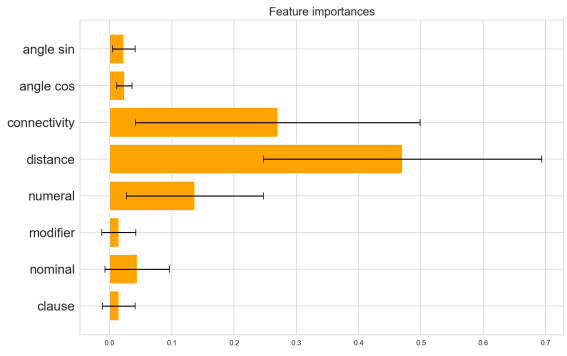


Figure 1:
The importance of each input feature, averaged over ten decision trees. The bars show standard deviation for each feature.

Our results suggest that layout and diagrammatic elements such as arrows and lines are crucial for making inferences about the multimodal discourse structure of diagrams. Detecting textual elements and lines may thus help to unpack the structure of diagrams. This has broader implications to emerging work in digital humanities, particularly within the paradigm of "distant viewing" (Arnold and Tilton 2019) and the growing interest in applying computational methods to multimodal data (Wevers and Smits 2020; Smits and Kestemont 2021). Compared to purely linguistic data, computational treatment of multimodal data in digital humanities rarely addresses fundamental questions such as how to identify basic units of analysis and the discourse relations that hold between them. Understanding the structure of multimodal discourse is a prerequisite for performing more complex analyses that are now regularly pursued using linguistic data, such as tracking semantic shifts. Achieving a similar capability for multimodal data requires a deeper understanding of discourse structures within individual modes of communication, such as the diagrammatic semiotic mode, and how individual modes are combined in multimodal artefacts.

Bibliography

Arnold, T. and Tilton, L. (2019). Distant viewing: analyzing large visual corpora. Digital Scholarship in the Humanities 34(Supplement 1): i3–i16.

Bateman, J.A., Wildfeuer, J. and Hiippala, T. (2017). Multimodality: Foundations, Research and Analysis. De Gruyter: Berlin.

Bateman, J.A. (2020). The foundational role of discourse semantics beyond language. In Zappavigna, M.

& Dreyfus, S. (eds) Discourses of Hope and Reconciliation. On J. R. Martin's Contribution to Systemic Functional Linguistics. Bloomsbury: London, pp. 39–55.

Hiippala, T. and Bateman, J.A. (2021). Semiotically-grounded distant viewing of diagrams: insights from two multimodal corpora. Digital Scholarship in the Humanities. DOI: 10.1093/llc/fqab063/6374705

Hiippala, T., Alikhani, M., Haverinen, J. et al. (2021) AI2D-RST: a multimodal corpus of 1000 primary school science diagrams. Language Resources & Evaluation 55: 661–688.

Smits, T. and Kestemont, M. (2021). Towards multimodal computational humanities: using CLIP to analyze late-nineteenth century magic lantern slides. In Proceedings of the Computational Humanities Research Conference (CHR), pp. 149–158.

Watanabe, Y. and Nagao, M. (1998). Diagram understanding using integration of layout information and textual information. In Proceedings of the 17th International Conference on Computational Linguistics (COLING), pp. 1374–1380.

When context matters. How to explore a knowledge graph of heraldic communication and its contexts of use in medieval and early modern Europe with methods such as graph embedding

Hiltmann, Torsten
torsten.hiltmann@hu-berlin.de
Humboldt-Universität zu Berlin

Schneider, Philipp
philipp.schneider.1@hu-berlin.de
Humboldt-Universität zu Berlin

The paper reports on the ongoing development of (1) a new Knowledge Graph for visual historical sources and their contextualisation, based on the *Digital Heraldry Ontology*, and (2) demonstrates by two examples how we approach the question of exploring these data to answer specific historical research questions.

The ontology focuses on the description of heraldic sources from medieval and early modern Europe. During this period, heraldry was omnipresent and constituted an important but still largely understudied instrument of visual communication used by vast parts of pre-modern European societies (Hiltmann 2019; 2018). It consisted of a mostly fixed set of symbolic primitives, that were combined into an abstract code of shapes and colors which were represented as coats of arms. Applied to supports as diverse as books, walls, stained glass and objects of all kinds, they could convey identity and status, property and claims, kinship or even abstract political and theological concepts – depending on the context in which they were used and by whom (Hablot 2019).

Therefore, in order to study them, context matters. This covers their materiality (where they have been represented), relevant actors (which person or institutional entity is represented by the coat of arms and who put it on) and the social role of those actors (in what function and for what purpose did they use the coat of arms). This complexity is modelled by the *Digital Heraldry Ontology*, encompassing the symbolic primitives of the coats of arms as well as the material and societal context of their usage. Drawing on multiple sources of different provenance, the project is building a Knowledge Graph that combines information on over 40,000 different coats of arms.

Using Semantic Web Technologies to deal with complex and heterogeneous data like this can be considered a standard in the fields of Digital History and Digital Cultural Heritage as an increasing number of large scale projects show (Dijkshoorn et al. 2018; Gehrke et al. 2015; Zamborlini and Betti 2017; Wang et al. 2020). However, this technology is, at least in historical research, only very rarely applied to generate new knowledge with data-driven methods, e.g. as early proposed by (Lin, Hong, and Doerr 2008). After briefly introducing the Knowledge Graph and its ontology, we will therefore focus on two examples and discuss ways how this data can not only be modeled but also explored for historical research.

First, we will show how querying and analyzing the ontology can be used to trace the development of coats of arms into increasingly complex means of communication and how this development can be placed in time and space. On the one hand, this is an important historical question that has not yet been answered (Hiltmann 2019), on the other hand, it allows us to discuss how ontologies can be applied to the study of specific research questions.

The second example shows how we use Graph Embeddings (Ristoski et al. 2019; Wang et al. 2017; Yang et al. 2020; El-Hajj et al. 2021) to transform the data into vector space and then to cluster murals and painted ceilings which feature coats of arms as means of communication. We will show how Graph Embeddings allow us to account for historical context in a scalable way, including the functional, temporal, and territorial context of the edifices in which the murals were displayed, as well as their patrons, to the extent that this information is available – and thus discuss the necessities for a successful application of this method.

Studying how to query over contextual data, we also address an open research problem for the analysis of Knowledge Graphs in general (Hogan et al. 2021).

To experts outside the domain of historical research, the talk will point out starting points to utilize Knowledge Graphs not only for the modeling but also for the data driven analysis of cultural heritage. This way, we want to contribute to open a dialogue on how to make use of the increasing number of data, provided by Knowledge Graphs, for domain specific research, profiting from their structure and flexibility, while also ensuring that general methodological as well as epistemological implications and challenges (Pierazzo 2019) inherent to historical data, are dealt with.

Acknowledgements

The project is funded by the Volkswagen Foundation (VolkswagenStiftung) under the title "Die Performanz der Wappen (2). Zur Ausdifferenzierung der heraldischen Kommunikation im hohen und späten Mittelalter (12.-15. Jahrhundert)".

Bibliography

Dijkshoorn, Chris, Lizzy Jongma, Lora Aroyo, Jacco van Ossenbruggen, Guus Schreiber, Wesley ter Weele, and Jan Wielemaker. 2018. 'The Rijksmuseum Collection as Linked Data'. Semantic Web 9 (2): 221–30. https://doi.org/10.3233/SW-170257.

El-Hajj, Hassan, and Matteo Valleriani. 'CIDOC2VEC: Extracting Information from Atomized CIDOC-CRM Humanities Knowledge Graphs'. Information 12, no. 12 (2021): 503. https://doi.org/10.3390/info12120503.

Gehrke, Stefanie, Eduard Frunzeanu, Pauline Charbonier, and Marie Muffat. 2015. 'Biblissima's Prototype on Medieval Manuscript Illuminations and Their Context'. In Proceedings of the First International Workshop Semantic Web for Scientific Heritage at the 12th ESWC 2015 Conference, edited by Arnaud Zucker, Isabelle Draelants, Catherine Faron-Zucker, and Monnin, Alexandre, 43–48. CEUR Workshop Proceedings. Portorož, Slovenia.

Hablot, Laurent. 2019. *Manuel de Héraldique Emblématique Médiévale*. Tours.

Hiltmann, Torsten. 2018. 'Arms and Arts in the Middle Ages. In *Heraldic Artists and Painters*, edited by Torsten Hiltmann and Laurent Hablot, 1:11–23.

———. 2019. 'Zwischen Grundwissenschaft, Kulturgeschichte Und Digitalen Methoden'. *Archiv Für Diplomatik* 65: 287–319.

Hogan, Aidan et al. 2021. 'Knowledge Graphs'. *ACM Computing Surveys* 54(4): 71:1-71:37. https://doi.org/10.1145/3447772.

Lin, Chia-Hung, Jen-Shin Hong, and Martin Doerr. 2008. 'Issues in an Inference Platform for Generating Deductive Knowledge'. *International Journal on Digital Libraries* 8 (2): 115–132. https://doi.org/10.1007/s00799-008-0034-0.

Pierazzo, Elena. 2019. 'How Subjective Is Your Model?' In *The Shape of Data in the Digital Humanities*, edited by Julia Flanders and Fotis Jannidis, 117–32. Routledge https://www.taylorfrancis.com/books/9781315552941.

Ristoski, Petar et al. 2019. 'RDF2Vec: RDF Graph Embeddings and Their Applications'. *Semantic Web* 10(4): 721–52. https://doi.org/10.3233/SW-180317.

Wang, Jun, Xiaoyu Li, Enhua Bian, Linxu Wang, Shuran Liu, and Nuo Chen. 2020. 'A Visualization-Assisted Reading Systemfor a Neo-Confucian Canon'. In DH2020. Book of Abstracts. Ottowa. https://dh2020.adho.org/wp-content/uploads/2020/07/729_AVisualizationAssistedReadingSystemforaNeoConfucianCanon.html.

Wang, Quan, et al. 2017. 'Knowledge Graph Embedding'. *IEEE Transactions on Knowledge and Data Engineering* 29(12): 2724–43. https://doi.org/10.1109/TKDE.2017.2754499.

Yang, Luwei et al. 2020. 'Dynamic Heterogeneous Graph Embedding Using Hierarchical Attentions'. In *Advances in Information Retrieval*, edited by Joemon M. Jose et al., 425–32. https://doi.org/10.1007/978-3-030-45442-5_53.

Zamborlini, Veruska, and Arianna Betti. 2017. 'Toward a Core Conceptual Model for (Im)Material Cultural Heritage in the Golden Agents Project'. In Workshops of SEMANTiCS 2017: Joint Proceedings of SEMANTiCS 2017 Workshops, Co-Located with the 13th International Conference on Semantic Systems (SEMANTiCS 2017), 4. Amsterdam.

Topic Modeling the Nineteenth-Century Poetry Canon

English Poetry Reprinted in Anthologies

Houston, Natalie
Natalie_Houston@uml.edu
University of Massachusetts Lowell, United States of America

The contents of poetry anthologies offer scholars a valuable resource for analyzing changes to the literary canon over time. No matter their size, poetry anthologies are necessarily selective, reprinting texts according to the editor's aesthetic, educational, or political decisions. Anthologies designed for use as textbooks describe

and define the field of literary study by providing a representation of a time period or literary movement within their pages. Within the academic context, these choices can have a far-reaching impact, as Wendell Harris suggests: "what is easily available in print tends to be what is being taught and written about" (Harris, 1991: 114). Anthologies also offer us a view into changing literary tastes and values: the poems by William Wordsworth (or any other poet) selected by anthology editors in the 1880s are very different from those selected by editors in the 1980s. As John Guillory suggests, "Canonicity is not a property of the work itself but of its transmission, its relation to other works in a collocation of works." (Guillory, 1993: 55) Poems accrue status and cultural value through being reprinted in anthologies, where they are placed in relation to other poems. This paper applies topic modeling and network analysis to a corpus of nineteenth-century British poems reprinted in British and American anthologies from 1880-2010 to understand the impact of those relationships.

Previous work used network analysis to examine the relationships among poets, poems, and anthologies in a corpus of 30 anthologies of nineteenth-century poetry published between 1880-2010 (Houston, 2017). A bimodal affiliate network of anthologies and poems reveals the relationships among anthologies that printed the same poems. A co-printing network (based on bibliographic co-citation analysis) consists of nodes representing each poem, with edges drawn between poems printed within the same anthology. Modularity analysis of this network reveals clusters of poems that are frequently printed together.

In this paper, I apply Latent Dirichlet Allocation (LDA) topic modeling (Blei et al) to the poems in the corpus and then examine the distribution of topics within the poem-anthology network and the co-printing network. LDA is a generative statistical model which assumes documents consist of "topics" made up of co-occurring words, and that these topics are present to varying proportions within the documents in the corpus. LDA has been shown to be an effective method for information retrieval and document classification tasks and has been applied to distant reading projects in the digital humanities on diverse materials ranging from novels to newspapers to scholarly articles (Buurma, 2015; Block, 2006; Goldstone and Underwood, 2012). Although the compressed semantic representation of an LDA topic can be seen as limiting the figurative complexity of poetic language (Rhody, 2012), the method has been shown to be effective for exploring and classifying short poetic texts (Navarro-Columbo, 2018; Plecháč and Haider, 2020; Šeļa et al, 2020).

Following Šeļa et al, I use an LDA topic model of the entire corpus as a representation of its semantic "topic space," an "abstracted representation of poetic language" (Šeļa et al, 2020: 15). Each poem can then be labeled by its highest-ranking topic (by proportion within the document). Encoding this semantic information as node features within the poem-anthology network reveals how the selections within particular anthologies emphasize or minimize particular themes. Within the co-printing network, this semantic information reveals how strongly thematic connections relate to the structural relationships of the poems' publication format. Combining the semantic insights offered by LDA topic modeling with the structural insights offered by network analysis offers new approaches to understanding the impact of influential anthologies (and their editors) in shaping subsequent generations' understanding of nineteenth-century British poetry.

Bibliography

Blei, D et al. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3: 993–1022.

Block, S. (2006). Doing More with Digitization. *Common-place* 6(2). http://commonplace.online/article/doing-more-with-digitization/.

Buurma, R. (2015). The fictionality of topic modeling: Machine reading Anthony Trollope's Barsetshire series. *Big Data & Society* 2(2): 1-6.

Goldstone, A. and Underwood, T. (2012). What Can Topic Models of PMLA Teach Us About the History of Literary Scholarship? *Journal of Digital Humanities* 2(1). http://journalofdigitalhumanities.org/2-1/.

Guillory, J. (1993). *Cultural Capital: The Problem of Literary Canon Formation.* Chicago and London: The University of Chicago Press.

Harris, W. (1991). Canonicity. *PMLA* 106: 110-21.

Houston, N.M. (2017). Measuring Canonicity: a Network Analysis Approach to Poetry Anthologies. *Digital Humanities 2017*. https://dh2017.adho.org/abstracts/479/479.pdf.

Navarro-Colorado, B. (2018). On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of Spanish Poetry. *Frontiers in Digital Humanities* 5(15). doi: 10.3389/fdigh.2018.00015.

Plecháč, P. and Haider, T. (2020). Mapping Topic Evolution Across Poetic Traditions. *Digital Humanities 2020*. https://dh2020.adho.org/wp-content/uploads/2020/07/600_MappingTopicEvolutionAcrossPoeticTraditions.html.

Rhody, L. (2012). Topic Modeling and Figurative Language. *Journal of Digital Humanities* 2(1). http://journalofdigitalhumanities.org/2-1/.

Šeļa, A. et al. (2020). Weak Genres: Modeling Association Between Poetic Meter and Meaning in Russian Poetry. *CHR 2020: Workshop on Computational Humanities Research*. http://ceur-ws.org/Vol-2723/long35.pdf.

Bert-based Chinese Buddhist Cannon Citation Extraction Model Utilizing Prior Defined Regex Pattern and Data Augmentation

HUNG, JEN JOU
jenjou.hung@dila.edu.tw
Dharma Drum Institute of Liberal Arts, Taiwan

Wang, Yu-Chun
ycwang@dila.edu.tw
Dharma Drum Institute of Liberal Arts, Taiwan

After Buddhism was introduced to China around the first century, monks and scholars interested in Buddhism began translating Buddhist scriptures. This translation activity produced an enormous number of Chinese Buddhist texts that were later compiled into the Chinese Buddhist Tripitaka. In recent decades, several Institutes have engaged in the Tripitaka digitization work. One of the significant achievements is made by CBETA Chinese Electronic Tripitaka Collection. It contains not only the electronic full text of the Tripitaka but also many related texts that are often needed for Buddhist studies. Because the Tripitaka has been properly preserved and well digitized, it has become the primary research material for studying Chinese Buddhism, and most Buddhist research papers quoted a lot of Tripitaka materials. From these citations, we can understand the relationship between the research papers and the Tripitaka texts. Furthermore, if we summarize and analyze the citation information, we may be able to understand the preference of Buddhist studies in citing Buddhist texts, which is very helpful for identifying the targets and trends of Buddhist studies.

The major difficulty encountered when extracting citations in the paper is the inconsistency and lacking of standard citation styles. Although most of the Tripitaka citations included almost the same information, the citation styles could be very different. For example: in the citation "T07, no. 220, p. 1068, c12", T07 represents the 7th volume of the Taisho Tripitaka, p. 1068 means page 1068, c12 denotes line 12 of column c, and the content at this location belongs to the text No. 220, the "Mahā-prajñāpāramitā Sūtra." In another example: the citation "T.1969, 47:170a21" means the citation is from the 21st line in column A on page 170 in volume 47 of Taisho Tripitaka. It can be seen that the information expressed by the two citations is similar, but the formats are quite different. The situation becomes even more complicated when we consider the citation with Chinese characters.

Our study has written a program using regex match to detect the citation strings and extract the reference location information. The performance of the detection program is quite good. So far, we have processed 3 Chinese Buddhist journals with 677 papers, and 21,029 valid Tripitaka citations have been extracted. The overall accuracy rate reaches 90.63%, and the recall rate is 94.68%. However, since we cannot obtain all possible reference formats initially, we can only design the corresponding regex patterns according to the reference instance encountered during the processing. This problem has led to a rapid increase in the number of regex patterns in our detection program. We already have created more than 70 patterns, and citations in tricky style still happen frequently. That drives us to seek a better strategy than endlessly adding regex patterns to our program.

The rapid development of modern artificial intelligence technology has brought us new ways to solve this problem. We plan to use the Joint Model based on Bidirectional Encoder Representations from Transformers (BERT) to establish the extraction Model of this research. Each sentence in the paper text will be used as the input of the BERT model, and an extra CLS token will be inserted at the front end of each sentence. Each token in the sentence will undergo multi-layer Transformers to generate the corresponding output vector. The output vector of CLS is connected to a fully connected binary neural network classifier for determining whether the sentence contains Tripitaka citation. The output of each subsequent token is also connected to a multi-class classifier to perform sequence labeling to determine which slot the token belongs to.

The deep learning models usually have a vast amount of parameters to be trained. If the amount of training data is insufficient, over-fitting problems may occur. Now, we only have the 677 marked full-text from Buddhist research papers to be used for training, but it seems insufficient in number. Therefore, we plan to use Text Data Augmentation to expand the training data set. Data Augmentation refers to adding noise to the training data to simulate more data to train the AI model. Such a technique is commonly used in training AI models for image recognition. However, in AI tasks based on text data, it is usually difficult to generate artificial training data because the rationality of the text content needs to be considered. However, we can overcome this problem by using more than 70 regex patterns to generate valid training data in our detection program. By combining the BERT mode and Text Data Augmentation technique, we expect to construct an effective Tripitaka citation detection and extraction mechanism.

Low-stakes activities for text analysis instruction in the undergraduate classroom

Isuster, Marcela Y.
marcela.isuster@mcgill.ca
McGill University

The introduction of digital text analysis tools and methodologies in (non-digital) humanities undergraduate courses has been sparsely documented in the literature. Furthermore, most of the times we encounter it, it is done in the context of semester-long or mid-term projects (Boyle and Hall 2016; Ficke 2014), where the stakes for the students are very high. Other times, they include a session on text analysis but no practical application of the tools and methodologies discussed in the course, other than a follow along demonstration.

This short paper introduces a middle point between these two extremes through the introduction of low stakes activities and assignments to help student discover and use digital text analysis tools and methodologies.

Besides giving students the opportunity to interact with the material in a safe and relaxed manner, low stakes activities help with student retention, confidence, and relationship building (Hamilton 2020; Meer and Chapman 2014). Low stakes activities are also a useful tool to assess comprehension and instruction when the person delivering the lesson is not the regular or official instructor in the course, such as the case of a librarian or a guest speaker. Furthermore, these types of activities are particularly useful for digital humanities instruction because they contribute to scaffolding, a method that has been identified as ideal in this type of instruction (Griffin and Taylor 2017; Isuster 2020; Sample and Schrum 2013; Tracy and Hoiem, 2018).

In the context of a Hispanic Studies course, a librarian offered a workshop series on digital text analysis and the web-based reading and analysis environment Voyant Tools. Interspersed with instruction there were a series of low stakes assessments that helped students understand and apply the content of the workshops. Working with the class readings, the librarian created activities that did not rely on having a single answer but encouraged students to discuss and interrogate both the methods and the information used. For example, when preparing a text for text analysis, students debated how different research questions necessitate different text preparation. The activities were completed in groups and were not graded. Results were discussed within the class.

The short paper presentation will explore the process of creating and implementing low stakes activities for digital text analysis and other digital humanities instruction. It will discuss the benefits of these types of activities as they pertain to digital humanities instruction and engagement and will share best practices and tips to help attendees create these kinds of activities in their own classrooms, including assignment design and sourcing materials.

Bibliography

Boyle, M. and Hall, C. (2016) 'Teaching "Don Quixote" in the Digital Age: Page and Screen, Visual and Tactile', *Hispania*, 99(4), pp. 600–614.

Ficke, S.H. (2014) 'From Text to Tags: The Digital Humanities in an Introductory Literature Course', *CEA Critic*, 76(2), pp. 200–210. 10.1353/cea.2014.0012.

Griffin, M. and Taylor, T.I. (2017) 'Shifting expectations: Revisiting core concepts of academic librarianship in undergraduate classes with a digital humanities focus', *College & Undergraduate Libraries*, 24(2–4), pp. 452–466. 10.1080/10691316.2017.1325346.

Hamilton, M. (2020) 'Implementation of a low-stakes daily assessment in a large introductory LAC course', *Teaching and Assessment Symposium* [Preprint]. Available at: https://digscholarship.unco.edu/posters_2020/4.

Isuster, M.Y. (2020) 'From students to authors: Fostering student content creation with Scalar', *College & Undergraduate Libraries*, 27(2-4), pp. 133–148. 10.1080/10691316.2020.1830908.

Meer, N.M. and Chapman, A. (2014) 'Assessment for confidence: Exploring the impact that low-stakes assessment design has on student retention', *The International Journal of Management Education*, 12(2), pp. 186–192. 10.1016/j.ijme.2014.01.003.

Sample, M. and Schrum, K. (2013) 'What's Wrong with Writing Essays: A Conversation', in Cohen, D.J. and Scheinfedlt, J.T. (eds) *Hacking the academy : new approaches to scholarship and teaching from digital humanities*. Ann Arbor, MI: University of Michigan Press, pp. 87–96.

Tracy, D.G. and Hoiem, E.M. (2018) 'Scaffolding and Play Approaches to Digital Humanities Pedagogy: Assessment and Iteration in Topically-Driven Courses', *Digital Humanities Quarterly*, 11(4). Available at: http://digitalhumanities.org:8081/dhq/vol/11/4/000358/000358.html.

The Inevitability of a Reproducibility Crisis in the Digital Humanities

Isuster, Marcela Y.
marcela.isuster@mcgill.ca
McGill University

Rod, Alisa B.
alisa.rod@mcgill.ca
McGill University

Once the purview of conventional science, sophisticated computational and algorithmic modeling methods are becoming ubiquitous approaches for digital humanists. However, unlike scientific approaches to research, digital humanities as a field lacks parallel paradigms or a research integrity framework related to reproducibility, also referred to as "replicability" in certain disciplines. Inherently, there is tension between humanistic approaches to inquiry and scientific or computational methods. However, as the methods and approaches of digital humanists increasingly mirror the methods and approaches of scientists in dealing with research data, it is becoming increasingly necessary to address the question of whether and how to apply scientific research integrity principles to digital humanities.

Peels and Bouter (2018) claim replication in the humanities is possible and desirable because research in these fields follows an epistemic process that, while discipline-specific, can be replicated. Because the digital humanities create knowledge through the use of algorithms and digital tools, the possibility and desirability of replication is even higher. Replication can be facilitated through the sharing of data and software that underlie research projects (Sikk, 2020). Unfortunately, the activity of sharing research materials is rare among both humanists more broadly and digital humanists more specifically.

To complicate matters, the rise in availability of third-party licensed text mining datasets from digital publishing vendors, library databases, and content providers, is enabling access to previously unavailable corpora and collections to analyze. Working with their institutional libraries, scholars may be able to access the text mining files for materials in their collections through platforms like the TDM Studio or the Gale Scholar Lab, they may purchase the files for specific publications, or in some cases they may simply request the data free of charge as part of the library's contract with the content provider. Unfortunately, these third-party licensed datasets are often subject to stringent legal terms and conditions. The potential inability of researchers to legally share the underlying data of their publication(s) presents a challenge regarding research transparency.

Through a brief overview of institutional case studies, this short presentation will explore the developing discourse surrounding reproducibility of digital humanities research and the subsequent inevitability of a digital humanities replication crisis contextually instigated by neoliberal constraints.

Bibliography

Peels, R., and Bouter, L. (2018). The possibility and desirability of replication in the humanities. *Palgrave Communications*, *4* (1): 1–4. https://doi.org/10.1057/s41599-018-0149-x

Sikk, K. (2020). Towards reproducible science in the digital humanities . *Digital History & Hermeneutics*. Retrieved from https://dhh.uni.lu/2020/05/19/towards-reproducible-science-in-the-digital-humanities-how-to-publish-your-data-and-code-alongside-your-research-with-the-help-of-zenodo/

The Linked Editorial Academic Framework: Creating an editorial environment for collaborative scholarship and publication

Jakacki, Diane Katherine
dkj004@bucknell.edu
Bucknell University, United States of America

Brown, Susan
sbrown@uoguelph.ca
University of Guelph, Canada

Cummings, James
james.cummings@newcastle.ac.uk
Newcastle University, United Kingdom

Ilovan, Mihaela
ilovan@ualberta.ca
University of Alberta, Canada

Black, Carolyn
carolyn@carolynblack.ca

Independent scholar

This short paper introduces LEAF (the Linked Editorial Academic Framework virtual research environment), an enhanced and expanded collaborative editorial platform that supports a variety of digital scholarly projects through a pipeline of integrated tools for collaborative production and publication of scholarly and documentary collections. Funded through the Canada Foundation for Innovation and the Andrew W. Mellon Foundation, LEAF aims to address the challenges that face many who undertake and maintain large-scale collaborative DH projects now: namely, the need to ensure that these projects can remain operational and available to editors and audiences over the long-haul. It is only by sharing physical, software, and human infrastructures across institutions that this can be accomplished. In so doing we can support scalability, interoperability, and preservation while allowing for dynamic, iterative, and collaborative editing, and therefore ensure that our materials, collections, and editions will remain viable and accessible. The LEAF team aims to do this by integrating best practices for text encoding, annotation, and metadata standards. This short paper will report on the development of LEAF and the functionalities that it will provide.

The implementation, and dissemination of LEAF is built upon a collaboration to extend the Canadian Writing Research Collaboratory (CWRC) built by the Universities of Alberta and Guelph (Susan Brown) with Bucknell University (Diane Jakacki), and Newcastle University (James Cummings) as founding partners. This work enhances CWRC's functionality through collaborative software development that will ultimately support multiple instances of the LEAF platform in Canada, the US, and the UK. At Bucknell, this work will inform the Liberal Arts Based Digital Edition Publishing Cooperative and the Bucknell Digital Press, funded by an Andrew W. Mellon Digital Publishing Cooperative Implementation grant that will support an expanding portfolio of peer-reviewed digital editions and edition clusters.

The LEAF platform combines hardware, software, and personnel. LEAF is being built on a solid foundation in terms of its data models, core functionality, and code management, so that it is positioned for extension and long-term sustainability. The platform is based on the Islandora 8 framework, which combines Drupal 8 with a Fedora 5 repository for long-term preservation. The LEAF repository will customize and enhance Islandora to enable digital humanities workflows and publication needs. Enhancements include an innovative web-based editing tool that allows users to employ TEI XML along with Web Annotation and IIIF standards-compatible Linked Open Data annotations that enhance discoverability and interoperability.

The founding LEAF institutions are collaborating to upgrade the existing CWRC environment and produce a fully modular platform that will also be hosted on Bucknell's servers, further tested at Newcastle University, and offered as containerized open-source code freely available for download and installation by other institutions. In particular, LEAF will facilitate the production and publication of dynamic digital scholarly editions and collections, offering multilingual transcription, translation, and image markup. Entirely browser-based, its functionality includes an in-browser XML markup editor, XML rendering tools, built-in text and data visualization tools including the Voyant Tools suite and its Dynamic Table of Contexts Browser. Overall the LEAF platform will provide a sophisticated interface for digital editions in which the XML markup is leveraged for navigation and active reading, and enhanced with Linked Open Data.

Document similarity and topic clues. A historiographical study case

Jolivet, Vincent
vincent.jolivet@chartes.psl.eu
École des chartes, France

Torres, Sergio
sergio.torres@chartes.psl.eu
École des chartes, France

Our University has recently published the longform abstracts of some 3000 institutional theses defended in history since 1849. The corpus is a rich documentary resource for historiographical studies. Unfortunately, there is no standard keyword indexing to browse this large collection and provide the reader with direct access to documents on the same subject. Such a functionality needs specific methods combining keywords, persons, places and in general intergroup patterns whose identification helps determine covered topics and related abstracts across more than a century. For this purpose, the proven clustering methods based on inter-document similarity are very effective, but in practice the interpretation of the similarity scores is difficult: a score describes how similar two documents are, but does not describe why they are similar. We have therefore experimented with methods combining document similarity and keyword extraction, so as to provide the researcher, in addition to a similarity score, with lexical clues facilitating the semantic interpretation of measured similarity.

In this presentation we present a pipeline leading with the extraction and formalization of indexing information in order to activate a document-similarity research engine, the evaluation of the scores obtained, as well as the benefits for information retrieval.

Methods

As our corpus is quite large, we preferred unsupervised approaches over supervised. The method is based on a semantic relatedness calculation using vectors, and the pipeline is composed of three steps.

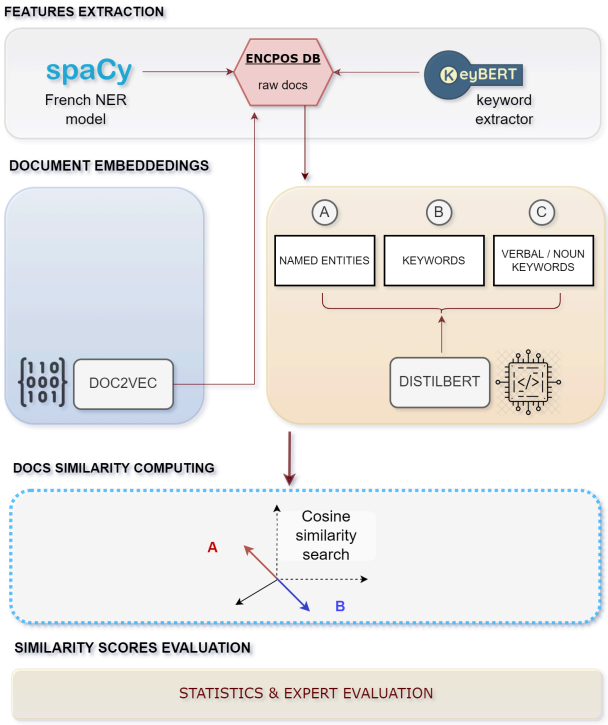


Figure 1:
Three steps for document similarity computing

1. We extract lexical and semantic features: (a) named entities (names, persons and organizations), using a French Spacy (Honnibal, M., Montani, I., 2017: 411) named entity extraction model based on CamemBert (Martin et al., 2019) language-model; and (b) keywords describing each abstract at a section level using KeyBert (Grootendorst, M., 2020), a keyword extractor based on the multilingual DistilBert (Sanh, V. et al., 2019) sentence embeddings library. To do so we apply embedding functions to our texts, mapping raw input data to low-dimensional vector representations. We then calculate the vector distance between the full-text embedding and candidate features embeddings to find the *top-k* candidates (the keywords) that are closest to the full text.

2. Our abstracts are then pre-processed into three versions containing: (a) their named entities, supposing that texts with the same entities are similar at a spatial and chronological level; (b) the keywords extracted during the first step at a paragraph-level and in so doing accounting for inflection variations such as tense and or stylistic elements; and (c) the only verbal and noun keywords, which keep only the phrasal root units to avoid lexical similarities and to summarize the text to its core components. Each one of these representations are later vectorized using Doc2vec (Le, Q., Mikolov, T., 2014: 1188), which generates context-independent embeddings (i.e, it collapses different word-meanings into a single vector), and DistilBert, which leverages Bert (Devlin, J. et al., 2018) to generate context-dependent sentence-level embeddings.

3. Finally, the cosine distance is calculated between the target-text and the database texts expressed as vectors to measure the document similarity score. This is obviously useful, as the score helps to identify related abstracts. Nevertheless, the similarity score doesn't provide all the necessary clues to determine the real performance of the given ranking of documents, and therefore must be evaluated further.

Evaluation

To estimate the relevance of the keywords embeddings method we calculate the similarity scores for all the documents pairs in an all vs all scheme using this method vs the Doc2vec and the Distilbert embeddings methods applied on full texts.

	key-distilbert				Distilbert			
	mean	median	var	stdev	mean	median	var	stdev
doc2vec	0.18	0.15	0.02	0.13	0.16	0.14	0.01	0.12
Distilbert	0.09	0.08	0.01	0.07	-	-	-	-

Figure 2:
Cosine distance statistics in an all vs all (≈ 9M matrix) scheme comparing three embedding methods: our key-distilbert method using the keywords (81 words on average) vs doc2vec and distilbert using the entire document (2013 words on average). var: variance, stdev: standard deviation

The statistics (median, mean, variance, standard deviation) indicate that in general the keyword approach, using 25x less amount of text, generates a similarity score very close (± 0.08 - 0.15) to the ones obtained using the full text (see Figure 2) on both methods, also proposing a time calculation 5x faster. This confirmation opens perspectives for the processing of very large corpora insofar as for close similarity scores.

Our method has another advantage, which was our initial goal: we provide for each pair of abstracts, in addition to a similarity score, a lexicon of the shared features (keywords and/or named entities) that we believe is useful for interpreting the score. For example, for two given abstracts, their high similarity score of 0.83 is enhanced by the lexicon of shared keywords "library", "manuscript", and "abbey"; these shared features seem to be clues for the semantic interpretation of the similarity. Not all cases are so obvious and the question of the relevance of these lexical clues for interpreting thematic similarities between documents is strongly raised.

Additionally, to evaluate the similarity scores as well as the relevance of the features lexicons, we submitted similar documents to experts, asking them to assess their degree of similarity and to rate the relevance of these shared lexicons to describe the link between the documents. This evaluation is ongoing.

Conclusion

Our initial objective was to obtain a reliable measure of the thematic similarity of the abstracts, by providing lexical clues useful for the semantic interpretation of the scores. We are convinced of the effectiveness of the method for the exploration of serial corpora such as cartularies or correspondences. Finally, the data produced are valuable for historiographical study, making it possible to quantify the most and least studied subjects diachronically, in particular through the analysis of the most associated keyword groupings.

Bibliography

Honnibal, M. and Montani, I. (2017). Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *Unpublished software application*.

Martin, L. et al. (2019). Camembert: a tasty french language model. arXiv preprint. arXiv:1911.03894.

Grootendorst, M. (2020). keyBERT: Minimal keyword extraction with bert.

Sanh, V. et al. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Le Quoc and Mikolov, T. (2014). Distributed representations of sentences and documents. *International conference on machine learning*. PMLR, 2014. p. 1188-1196.

Devlin, J. et al. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Information Platform for Linked Data of Regional Historical Materials and its Agent Name Finding Process

Kameda, Akihiro
cm3ak@outlook.com
National Museum of Japanese History, Japan

Goto, Makoto
m-goto@rekihaku.ac.jp
National Museum of Japanese History, Japan

Overview

This presentation describes the construction of a system and the analysis and maintenance of data for the advanced use of the inventory of regional historical resources, especially using interactive annotation of agent names. We are driving a project for the inheritance and preservation of regional historical resources. In order to achieve the objectives of this project, we have developed a data infrastructure for advanced use of the inventories of regional historical resources. In particular, we aimed to create a system in which computers help people to discover information, rather than the conventional system in which people search and browse directly. Specifically, we resolved orthographical variants and integrated values, constructed identifiers and URIs, and described the provenance and components of resources. As a result, we were able to provide a Linked Data for regional historical resources, and we found the design of appropriate information infrastructure and its data generation process. In regional historical resources, there are many people and companies described. Some people can be associated with clans and positioned in family tree diagrams, other people are nameless and their detailed profiles are unknown. Those agent names and relationships among them in the archive of regional historical resources characterize the archive itself. If we only focus on some famous people already known in other documents, it is efficient to bring the dictionary of those names including alternative names and find those names in the archive. However, some names which are quite frequently used in the archive and not so much known in other documents are worth being analyzed and described. So, we extract the candidate names from the archive, list up the information of famous people from external resources

such as encyclopedias of history, and extend the list as the archive specific directory based on frequency and co-occurrence analyses.

Platform: khirin-c

We have a series of systems named *khirin*(Knowledgebase of Historical Resources in Institutes). The system described in this presentation is named *khirin-c*, which collaborate with a IIIF based image presentation system *khirin-i*. *Khirin-c* stores four types of URIs: (1)dataset-wise, (2)type-wise, (3)accompanying information of type 1 , and (4)external URIs. We can import data from some format including the RDF 1.1 Turtle [1], which we most frequently use. Moreover, *khirin-c* has web-based GUI interface and we can edit the data online. To set the relationships among the dataset and options of HTML rendering, we use GUI interface mainly. To edit the data, we do it offline so that we can maintain the data and track the changes and versions.

Agent Name Finding Process

In this presentation, we focus on describing how we organized information about people and organizations to support their reading and understanding. First, we extract candidate names using Entity Names Recognition software GiNZA [2]. In the cleansing stage, for the data received in tabular form, we set tentative identifiers, clustered the values using OpenRefine [3], and used Wikidata to expand and identify the information. To support the deciphering of the data, we provided some interfaces which can narrowed down the list by each person or organization related to each item and visualized the network based on the sender-recipient relation of letters.

Notes

1. https://www.w3.org/TR/turtle/
2. https://github.com/megagonlabs/ginza
3. https://openrefine.org/

Processes and Practicalities in Developing and Sustaining a Text Mining Platform: Gale Digital Scholar Lab

Ketchley, Sarah
ketchley@uw.edu
University of Washington/Gale, United States of America

Ludwig, Jess
jess.ludwig@cenage.com
Gale

Gale Digital Scholar Lab was developed in 2018 to fulfil requests for a platform for text mining primary source documents without necessarily having to learn to code in Python or R. Based on beta-testing user interviews, it was determined that some of the most significant barriers to entry into the field of text-based digital humanities data mining include not knowing how or where to start in order to build a DH project, not having time to gather a significant data set or to clean and organize data for analysis, and having limited institutional infrastructure and support for projects that include text mining methodologies. In designing the DS Lab, the goal was to provide a scaffolded experience for users new to the field of digital humanities, while offering options for extensibility for researchers with established projects. This included providing pathways for research, teaching and learning by both students, faculty, and librarians.

The DS Lab has been iteratively developed since its first release, with updates including tool enhancements and support for pedagogical use of the platform, and more recently a platform migration, workflow tweaks and improved accessibility. The DS Lab integrates six GUI-based tools for conducting text analysis of primary source archives and user-uploaded plaintext documents. These tools comprise Named Entity Recognition, Sentiment Analysis, Ngrams, Parts of Speech Tagging, Topic Modeling and Clustering. Recognizing that quality of OCR text is key in achieving meaningful analysis outputs, the DS Lab also presents options for text cleaning as part of the curation process. Import and export of text and metadata are also supported.

To orient users who are new to the field of DH to the workflow and outcomes, the platform incorporates an extensive Learning Center with contextual help documentation including brief recorded videos, images, text, and sample projects. Similarly, for teachers who are looking for ideas or additional support in the platform, there are draft syllabi, outline learning objectives, and downloadable project outlines.

This 10-minute talk will focus on describing the process and challenges of developing the DS Lab interface, meeting the often-competing demands of balancing developer time, the scope of individual project sprints, and projected cost. Consideration will be given to the workflows which were successful as well as those that needed to be adapted or scrapped altogether. It will discuss using personas to design the features and functionality in the platform, and the advantages and drawbacks of doing so. The development of the Learning Centre is a case in point, since its development drew on a range of internal and external expertise such as academic advisors, curriculum developers,

and UX designers as well as in-house software and content engineers, metadata and content architects and the product and archives team. This collaborative undertaking took considerable management to balance expectations and outcomes, and to ensure that communication flowed clearly. These considerations are not unique to the Gale Digital Scholar Lab development project but can be extrapolated to other similar DH projects. The intent of the talk is to highlight how the lessons learned by the Gale development team and external stakeholders can be used to provide guidance for others.

Bibliography

Besette, Lee. (2012). "Challenges in Digital Humanities." *Inside Higher Ed.* https://www.insidehighered.com/blogs/college-ready-writing/challenges-digital-humanities. Accessed 16 March 2022.

Campese, C., Thiago Bertolini, d. S., Lorena Pereira, d. C., & Janaina Mascarenhas, H. C. (2019). *User stories method and assistive technology product development: A new approach to requirements elicitation*. Cambridge: Cambridge University Press. doi:http://dx.doi.org/10.1017/dsi.2019.385

Coutu, Diane. (2015). "Why Teams Don't Work." *Harvard Business Review.* https://hbr.org/2009/05/why-teams-dont-work. Accessed 25 Apr. 2022.

Currier, Brent D. (2017). "They Think all of this is new: Leveraging Librarians' Project Management Skills for the Digital Humanities." *College & Undergraduate Libraries* 24, 270-289. https://doi.org/10.1080/10691316.2017.1347541

Dingsøyr, Torgeir, et al. (2018). "Coordinating Knowledge Work in Multiteam Programs: Findings From a Large-Scale Agile Development Program." *Project Management Journal*, vol. 49, no. 6. 64–77, doi:10.1177/8756972818798980.

Gratton, Lynda, and Tamara J. Erickson. (2016). "Eight Ways to Build Collaborative Teams." *Harvard Business Review*. hbr.org/2007/11/eight-ways-to-build-collaborative-teams. Accessed 25 Apr. 2022.

Jenkins, Nick. (2008) *A Software Testing Primer An Introduction to Software Testing*. San Francisco: Creative Commons.

Nielsen, Jakob. (nd). "Why You Only Need to Test with 5 Users." *Nielsen Norman Group*. *www.nngroup.com*, https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/. Accessed 30 Apr. 2022.

Siemens, L. (2011). The Balance between On-line and In-person Interactions: Methods for the Development of Digital Humanities Collaboration. *Digital Studies/le Champ Numérique*, *2*(1). DOI: http://doi.org/10.16995/dscn.259. Accessed 16 March 2022.

After deplatforming: digital methods for documenting Twitter and YouTube moderation

De Keulenaar, Emillie Victoria
e.v.de.keulenaar@rug.nl
University of Groningen, OILab

Kisjes, Ivan
i.kisjes@uva.nl
University of Amsterdam

Following myriad controversies, including harassment campaigns and dissemination of conspiratorial narratives (Bounegru *et al.*, 2018; Jeong, 2019) , genocides (Mozur, 2018) and political extremism (Ganesh and Bright, 2020), social media platforms and cloud hosts have deployed additional measures to moderate user-generated contents. YouTube, Facebook and Twitter, in particular, have further developed content moderation techniques that prevent the circulation of "problematic information" (Jack, 2017) via deletion or "deplatforming" (Rogers, 2020), automatically flagging content as "misleading" (Gorwa, Binns and Katzenbach, 2020), and demoting or "shadow-banning" (Myers West, 2018) user posts across search and recommendation results (Goldman, 2019). To date, millions of content associated with hate speech, COVID-19 conspiracy theories and incitement to violence have been wiped out from Twitter (Al Jazeera, 2021), YouTube (Keulenaar, Burton and Kisjes, 2021), Instagram (Francis, 2021), and Facebook (Lerman, 2020).

Though it is essential for regulating any kind of public sphere, the imperative to delete and otherwise obfuscate problematic content has brought new challenges to new media and digital humanities scholars in at least three ways. Most directly, it renders problematic content unarchivable by default, making almost impossible the study of already precariously archived Web and social media data (Brügger and Schroeder, 2017). Second, moderated platforms provide little information on how they practice moderation. This prevents public efforts from scrutinizing the normative framework platforms adopt to determine what can and cannot be said. In turn, the obfuscation of moderated content and moderation practices further hampers studies on speech moderation and norms as a key historical and societal practices to any modern-day society. Speech norms join

a long history of legal, social and political measures to prevent the normalization of problematic histories, and have been implemented in the form of laws, social conventions and civil right campaigns in a range of media types.

Thus far, scholars have relied on a handful of improvised methods for studying moderation and moderated contents. With the exception of Twitter's Enterprise API (Twitter, 2021), platforms rarely outsource information about what specific contents they have moderated and why. While most studies in platform governance focus on written documentation and leaked documents from ex-platform employees (Wall Street Journal, 2021), digital methods research tends to rely on information provided by Application Program Interfaces (APIs). These researcher practices have yet to be systematically combined, and are still prey to changing APIs (Perriam, Birkbak and Freeman, 2020) and reprisals for illegal scraping of information that is otherwise not publicly provided(Bond, 2021).

On this matter, this paper discusses a new "digital forensics": a set of methods one can use to reconstruct platform and user traces. In other words, it proposes methods to reconstruct the scene after or on which platform or user data has disappeared in the context of one specific platform effect: content moderation. Drawing from two case studies on Twitter and YouTube's moderation of COVID-19 misinformation between January and April of 2020 (de Keulenaar *et al.*, Forthcoming), and hate speech on YouTube between 2007 and 2018 (de Keulenaar *et al.*, 2021), it proposes a web historiography (Brügger, 2013) of content moderation policies and techniques with a combination of HTML scraping, retrieval of moderation metadata via APIs, and detection of content availability through dynamic archival of moderated contents.

It describes the potentials and shortcomings of four methods:

1. **A contextualization of content moderation practices**. Using the Wayback Machine to trace changes in content moderation policies, this method consists in systematically annotating changes to (a) how the platform decides what is and is not problematic; and (b) observing and documenting the techniques the platform uses to moderate contents respectively.
2. **Dynamic archiving of content susceptible to being moderated**. This consists in first developing a taxonomy of problematic speech based on what platform content moderation policies consider to be problematic, such as hate speech (examples being speech that targets gender, racial and other identities), misinformation (such as information that contradicts COVID medical authorities) and "borderline content" (information likely to infringe upon policies in the future, usually described as conspiratorial or fringe discourses). This allows us to design queries that reflect problematic speech, which we use to collect corresponding tweets, videos and comments daily, as well as its corresponding metadata and platform effects (search and recommendation rankings, etc.).
3. **Reverse-engineering content moderation practices**. Using Twitter Academic and YouTube's standard APIs to reverse-engineer moderation practices, collecting metadata for: (1) the availability of problematic contents per day; (2) by-products of algorithmic moderation, such as their ranking in search and recommendation results over time, and flags and prompts we scrape using Selenium; and (3) user engagement in moderated contents.
4. **Tracing the effects of the disappearance of moderated data** in one platform by looking at its migration in other platforms. This implies looking at how users react to the practice of moderation (for example, what they say about "cancelling", "deplatforming", "deleting", "shadowbanning" and other forms of platform interventions), as well as how they curtail these sanctions by access sanctioned content in alternative or "alt-tech" platforms like Telegram, Bitchute and Parler.

Though imperfect, we aim to demonstrate how this ensemble of methods allows one to contextualize moderation practices, such as deplatforming, algorithmic demotion and flagging, within content moderation policies around hate speech and misinformation. We argue that they allow researchers to surface volatile content moderation practices, as well as map the larger effects of deplatforming across the Web in the fragmentation of users' information diets across a fringe-to-mainstream social media ecology. Most importantly, we propose this method as a way to systematically document online speech moderation practices and contribute to a history of speech norms across political contexts and media types.

Victorian400: Colorizing Victorian Illustrations

Kim, Hoyeol
elibooklover@gmail.com
Texas A&M University, United States of America

I introduce the Victorian400 dataset for colorizing black and white nineteenth century illustrations using deep learning. While there has been progress in colorizing photos and videos based on Conditional Generative Adversarial

Networks (cGANs, there have not been attempts to colorize illustrations from the nineteenth century using deep learning. In addition, datasets with nineteenth-century illustrations have not been provided for deep learning colorization. Therefore, I decided to create the Victorain400 dataset, which is a collection of colorful illustrations painted with nineteenth-century palettes. The Victorian400 dataset provides an opportunity for those studying deep learning to run code easily without high performance devices. I have created, curated and publicly shared the Victorian400 dataset for the deep learning colorization of illustrations from the Victorian era.

I examine the process of creating and curating the Victorian400 dataset and demonstrate the validity of the Victorian400 dataset by looking into the results of the test set with the trained Victorian400 set. I tested the Victorian400 dataset with the pix2pix model introduced by Isola et al., which performs automatic graphic operations on photographs based on cGANs by learning from datasets. It was built based on the GAN model which was introduced by Goodfellow et al. The GAN model has the generator and the discriminator learn and compete with each other for the generation of the best outputs. Similarly, the generator and the discriminator in cGANs can be used to predict colors for input images through the usage of encoders and decoders. The pix2pix model uses U-Net, a convolutional network for image segmentation with skip connections, to make it possible to get both local and contextual information quickly. After showing the test results with the test set from the Victorian400 dataset, I colorize black-and-white illustrations from Charles Dickens's Bleak House (serialized 1852–1853) for which the plates were created by Hablot Knight Browne (Phiz). Through the experiments, I reveal that the illustrations created with the dark plate technique are compatible with deep learning colorization due to the distinct contrast between darkness and highlights. I discuss the limits of colorizing black-and-white illustrations with the Victorian400 dataset, such as the lack of colors in backgrounds and the possible distortion of original illustrations. Ultimately, I claim that colorized illustrations provide imagination and enjoyment to modern readers when reading fiction.

The Victorian400 dataset was created for data scientists and digital humanists who create, train and test colorization deep learning models. The Victorian400 dataset will not only save a tremendous amount of time for digital humanists who experiment with Victorian illustrations, but will also contribute to the development of deep learning-based research in the digital humanities. As a digital humanist, I believe that we should create, curate, and share humanities datasets for deep learning like the Victorian400 dataset, as well as perform exploratory data analysis, since humanities datasets created by digital humanists are credible enough to be deployed for deep learning.

An experiment in agent-based probabilistic city population reconstruction

Kisjes, Ivan
i.kisjes@uva.nl
University of Amsterdam, Netherlands, The

Van Wissen, Leon
l.vanwissen@uva.nl
University of Amsterdam, Netherlands, The

Reconstructing past population configurations is no easy task, even when detailed demographic information exists such as birth, death and marriage records (e.g. Bailey et al. 2020, Efremova 2016). Reconstruction is necessary to be able to disambiguate and link people mentioned across various historic sources. The best way to do it is by hand, but that is prohibitively labor intensive. For automatic methods, name and spelling variations and uncertain relations between mentioned names present large problems (see e.g. Idrissou et al. 2018). Missing information does, too: demographic records tend not to include information on migrations into or out of the city in question, in our case Amsterdam.

Methods that are being tried tend to reconstruct individuals on the basis of detected name identifications and their relations to other mentioned names (e.g. Bloothooft et al 2015, Bailey et al. 2020). While this seems promising, it fails to include other information we may make use of: demographic and biological statistics (e.g. Störmer 2018, Alter and Clark 2010). Using the latter, we should be able to identify a person mentioned with one name in a certain source with the same person mentioned in another by only a nickname without being dependent on related names being mentioned as well.

We explore employing a temporal, iterative agent interaction model, similar to those used in biological population evolution models (e.g. Toni and Stumpf, 2010, Marchetti et al 2017,Levin et al 1997). We set up a Python actor model using MESA (Kazil et al. 2020) that can iterate over time and has access to all birth, death and marriage 'events' that are known from the archival records. We iterate over a 50-year sample in the dataset (for Amsterdam data is available from ca. 1600 through 1940, we use 1750-1800 as a test case).

Each year, all names mentioned in 'events' in that year become actors (in the model sense). So if there is a birth record, the name of the child will be 'born' in that year,

and the parents become actors that have children. The actors then decide how they ended up existing in that year: are they representations of actors that already exist in the current model because of a mention in a previous year? Are they new migrations into the city? Are they births of new citizens, or perhaps deaths of previously known or unknown actors? Do they (re)marry, should they have been married before that year? Actors make these decisions based on the event, their existing relations within the currently running model, and probability.

The decisions the actors make are based on statistics of the population in question combined with biological statistics and limitations. For example, it is very unlikely that people in the dataset live longer than 130 years, but very young children also die often, so mortality probability changes over an actor's lifespan. But we also know that mortality rates are not constant over the whole timespan – certain years may have seen disasters, increasing mortality (see e.g. Aberth 2013, Jensen 2019), and others may have seen immigration waves, increasing population size without a higher birth rate. Other basic probabilities also affect the model: e.g. women rarely have children before age 11 or after age 60, and unmarried women are more likely to be childless than married ones.

Running this model we end up with a temporally changing network that we can evaluate in different ways in order to approximate historic reality. First, we can use general demographic knowledge to evaluate it, for example known population sizes (e.g. Lourens and Lucassen 1997, Paping 2014, Frijhoff et al 2006 etc.), mortality rates, birth rate estimations etc. (e.g. Heathcote 2001). Secondly, we can use network properties (e.g. population spikiness, average number of children, average age etc.) to evaluate. A third way is to deduce how 'solid' each actor's life is, recording e.g. whether they have parents, whether the parents have the same last name as themselves, whether they have any relations to others in the model at all, whether they have only one or multiple mentions in the records etc. A fourth way is to compare them to known prosopographic data (we use ECARTICO). We test each of these methods and combine them into a meta-evaluation method.

Re-running the model many times, resulting in different models, we evaluate each in order to select the one best approximating reality and will discuss our findings.

Bibliography

Aberth, John (2013): *From the Brink of the Apocalypse: Confronting Famine, War, Plague and Death in the Later Middle Ages*. 2nd Edition 2013, First Published 2010

Alter, G., & Clark, G. (2010). *The demographic transition and human capital*. In S. Broadberry & K. O'Rourke: The Cambridge Economic History of Modern Europe, pp. 43-69. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511794834.004

Bayley, Martha J., Connor Cole, Morgan Henderson and Catherine Massey (2020): *How Well Do Automated Linking Methods Perform? Lessons from US Historical Data*. Journal of Economic Literature Vol.. 58, NO. 4, December 2020 (pp. 997-1044)

Bloothooft, Gerrit, Peter Christen, Kees Mandemakers and Marijn Schraagen (2015): *Population Reconstruction*. Springer

Efremova, I. (2016). *Mining social structures from genealogical data*. Technische Universiteit Eindhoven.

Frijhoff, W., M. Prak and M. Hel (2006): *Geschiedenis van Amsterdam, II-2, Zelfbewuste stadstaat 1650-1813.* Bijdragen en mededelingen betreffende de geschiedenis der Nederlanden 121(3):500 DOI:10.18352/bmgn-lchr.6475

Heathcote C., Higgins T. (2001) A Regression Model of Mortality, with Application to the Netherlands. In: Tabeau E., van den Berg Jeths A., Heathcote C. (eds) Forecasting Mortality in Developed Countries. European Studies of Population, vol 9. Springer, Dordrecht. https://doi.org/10.1007/0-306-47562-6_3

Idrissou, A., Zamborlini, V., Latronico, C., van Harmelen, F., & van den Heuvel, C. M. J. M. (2018). *Amsterdamers from the Golden Age to the Information Age via Lenticular Lenses*: Short paper.

Jensen, L.E. (2019): *'Disaster upon Disaster Inflicted on the Dutch'. Singing about Disasters in the Netherlands, 1600-1900*. Bijdragen en Mededelingen Betreffende de Geschiedenis der Nederlanden, 134, 2, (2019), pp. 45-70 https://doi.org/10.18352/bmgn-lchr.10449

Kazil, Jackie, David Masad, and Andrew Crooks (2020): *Utilizing Python for Agent-Based Modeling: The Mesa Framework*. In Robert Thomson, Halil Bisgin, Cristopher Dancy,Ayaz Hyder and Muhammad Hussain: Social, Cultural, and Behavioral Modeling. Springer International Publishing

Levin, Simon A., Bryan Grenfell, Alan Hastings, Alan S. Perelson (1997): *Mathematical and Computational Challenges in Population Biology and Ecosystems Science*. Science 1997 Vol 275, Issue 5298 pp. 334-343 DOI: 10.1126/science.275.5298.334

Lourens, Piet, and Jan Lucassen (1997): *Inwonertallen van Nederlandse steden ca. 1300-1800*. Amsterdam: Vereniging Het Nederlandsch Economisch-Historisch Archief..

Marchetti, Luca, Corrado Priami and Vo Hong Tanh (2017): *Simulation Algorithms for Computational Systems Biology*. 2017 Springer 331963111X

Newton, G., & Bennett, R. (2020). *Record-linkage of entrepreneurs in the England and Wales Censuses 1851-91 using BBCE and I-CeM*. https://doi.org/10.17863/CAM.50178

Paping, Richard (2014): *General Dutch Population development 1400-1850*. 1st ESHD conference, Italy

Störmer, Charlotte, Corry Gellatly, Anita Boele, and Tine De Moor (2017): *Long-Term Trends in Marriage Timing and the Impact of Migration, the Netherlands (1650-1899)*. Historical Life Course Studies 6 (December):40-68. https://doi.org/10.51964/hlcs9327.

Toni, Tine, Stumpf, Michael P.H (2010): *Simulation-based model selection for dynamical systems in systems and population biology.* Bioinformatics, Volume 26, Issue 1, 1 January 2010, Pages 104–110, https://doi.org/10.1093/bioinformatics/btp619.

Data Diffraction: A Counternarrative to Integration in Digital Humanities Research

Kleymann, Rabea

kleymann@zfl-berlin.org
Leibniz-Zentrum für Literatur- und Kulturforschung, Germany

Mixed methods are firmly established in digital humanities (DH) scholarship. While this approach is understood as a research design adopted from social sciences, mixed methods seem to be an umbrella term for defining DH's methodological framework in general (Sá Pereira 2019; Herrmann 2017). The use of computational procedures in DH is often regarded as a combination of quantitative and qualitative methods. Further conceptual pairs, for example *close* and *distant reading*, go hand in hand with this. However, mixed methods research rests on two premises (Uprichard and Dawney 2019, 20). [1] First, the research design indicates that the complexity of an epistemic object is addressed by the plurality of methods used (Fieldling 2012, 127). Second, research data obtained by mixed methods can be integrated. In other words, results of different methodological settings can be put into a coherent narrative. So far integration within mixed methods research is often discussed in a realm of technical challenges concerning data settings, standards and ontologies. Although data integration addresses epistemological and social issues of conformity and interoperability of research data for a global DH community.

In this short presentation, I argue that integration provides one device to explore questions of difference and diversification within DH scholarship. Therefore, I investigate promises, constraints and pitfalls of the "integration"-narrative, which seems to be deeply enfolded in mixed methods research. The focus of attention will be on compatibilities as well as forms of inferences, which gain relevance manufacturing of knowledge within mixed methods research (Knorr Cetina 1981; Kuhn 1994). In order to tackle these questions, I discuss "data diffraction" (Uprichard and Dawney 2019, 26) – a counternarrative presented by Uprichard and Dawney – as one complementary aim for dealing with different data settings resulting from mixed methods research. What new perspectives open up if we speak of data diffraction instead of data integration?

The term *diffraction*, which was originally introduced by Donna Haraway and Karen Barad for epistemological endeavors, initially describes optical interference patterns that arise when two waves are superimposed (Haraway 1992, 300; Barad 2007, 91). Contrary to an holistic idea of integrating parts under a whole, diffraction is about the productive maintenance of differences. Exploring the narrative of data diffraction, this short presentation brings into sharper relief latent integration mechanisms on the one hand, and explore possible alternatives on the other (Drucker 2021, 2; Liu 2020, 130). Beyond or complementary to integration, how could methods or data relate to each other? What if, we explicitly describe diffractions, that is incommensurabilities and dissonances, of methods and data? What would this mean for international collaboration?

Two examples are shortly discussed in this presentation. The first example brings into focus data integration in the context of mixed methods through ontologies as formal models. Ontologies enable to store and query mixed research data. Therefore, ontologies promise a semantic interoperability that allows different data sets to be integrated with each other (Pidd and Rogers, 2018). But how are different data settings handled? What possibilities do OWL and RDF schemas offer to describe leftovers and surplus of research data? In this context, I speculate about possibilities for data diffraction. One scenario here is ontology hijacking (Eide and Smith-Ore 2019, 188).

The second example dwells on existing mixed methods approaches from the literary studies, digital stylometry in particular. "DH style studies may be a natural environment for the mixed-methods-paradigm", as Herrmann has phrased it (Herrmann 2017). In digital stylometry, for instance, authorship attribution with Burrows' Delta algorithm, agglomerative cluster analysis as well as principal component analysis are widely used (Karsdorp et al. 2021, 248f.). Using the literary category of style, I

examine how integration and diffraction might differently enact and constitute style as an object of inquiry within mixed methods research. In doing so, I engage a critical reading of two python scripts from digital stylometry studies. Where does data integration or diffraction actually take place in concrete terms?

Bibliography

Barad, K. (2007). *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Durham: Duke Univ. Press.

Drucker, J. (2021). Viewpoint: Hetero-ontologies and taxonomies in the wild. *Art Libraries Journal* 46 (2), 36–39. https://doi.org/10.1017/alj.2021.2.

Eide, Ø. and Smith-Ore, C.E. (2019). Ontologies and data modeling. In Flanders, J. and Jannidis, F. (eds), *The shape of data in the digital humanities. Modeling texts and text-based resources*. London/New York, Routledge, pp.178–196.

Fielding, N. G. (2012). Triangulation and Mixed Methods Designs. *Journal of mixed methods research* 6 (2): 124–136. https://doi.org/10.1177/1558689812437101.

Haraway, D. (1992). The Promises of Monsters: A Regenerative Politics for Inappriopriate/d Others. In Grossberg, L., Nelson, C. and Treichler, P. A. (eds), *Cultural Studies*. New York, NY, London: Routledge, pp. 295–337.

Herrmann, B. (2017). In a Test Bed with Kafka. Introducing a Mixed-Method Approach to Digital Stylistics. *Digital Humanities Quarterly* 11, (4). http://www.digitalhumanities.org/dhq/vol/11/4/000341/000341.html.

Karsdorp, F., Kestemont, M. and Ridell, A. (2021). *Humanities Data Analysis: Case Studies with Python*. Princeton: Princeton University Press.

Knorr Cetina, K. (1981). *The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science*. 1. Edition. Oxford: Pergamon Press.

Kuhn, T. (1994). *The Structure of Scientific Revolutions*. Chicago: Chicago Univ. Press, 1994.

Liu, A. (2020). Toward a Diversity Stack: Digital Humanities and Diversity as Technical Problem. *PMLA* 135, (1): 130–51.

Pidd, M. and Rogers, K. (2018). Why Use an Ontology? Mixed Methods Produce Mixed Data. October 18, 2018, https://talkinghumanities.blogs.sas.ac.uk/2018/10/18/why-use-an-ontology-mixed-methods-produce-mixed-data/ (accessed 20 April 2022).

Sá Pereira, M. P. (2019). Mixed Methodological Digital Humanities. In Gold, M. K. and Klein, L. F. (eds), Debates in the Digital Humanities 2019. 5. Minneapolis: University of Minnesota Press, 2019.

Uprichard, E. and Dawney, L. (2019). Data Diffraction: Challenging Data Integration in Mixed Methods Research. Journal of mixed methods research 13, (1): 19–32. https://doi.org/10.1177/1558689816674650

Notes

1. This paper connects directly to Uprichard & Dawney's research approach, which already addresses the problem of integration as well as data diffraction in mixed-methods approaches in the social sciences.

Contabilizar el comercio imperial: Analysis of early double-entry accounting books using the TEI/DEPCHA

Kokaze, Naoki
xiao3feng10324@yahoo.co.jp
Chiba University, Japan

Fushimi, Takeshi
taquito@keio.jp
Keio University, Japan

Nakamura, Yusuke
y-nkmr@l.u-tokyo.ac.jp
The University of Tokyo, Japan

Introduction: The Purpose of Research

We have been analyzing accounting books in order to understand the economic history of the early Spanish empire. This presentation focuses on the double-entry accounting books of the Salamanca company in Burgos, a city in northern Spain, in the mid-sixteenth century. This paper reports our materials, analytical framework, progress, and prospect. In this research, data mining from the accounting book, structured with the Text Encoding Initiative /DEPCHA, has enabled historical insights into the unique terminology of accounting, the balance of payments, and the tendency of abbreviation.

Materials

Our materials are found among the accounting books in the Archive of Burgos Provincial Congress (Table 1). We picked out as an example a pair of accounting books based on the double-entry method (classified as CM32 and CM108). In this method, two types of books are required. One is the journal (diario in Spanish), in which all transactions are recorded chronologically. The other is the ledger (libro mayor), where each transaction in the journal must be transcribed twice, one on the left page as a debit transaction, and the other as a credit on the right (cf. Figure 1). This double transcription, a quintessential characteristic of the method, must be encoded with TEI.

Each transcription in the ledger has at least two sections: the detail of the transaction on the left; the value of the transaction on the right. On average, the left section contains information such as the name of the debtor, date and place of transaction, merchandise and its price and quantity, names of other parties involved, terms of the transaction, and page number of the corresponding transcription. As of March 2022, we have marked up 12 folios of the ledger, which contains 197 entries.

	Title	Period	Classified Number – CM+	Type of Accounting	DIGITIZED	Owner	Number of Folios	Size of Folio
1	Manuel (o Libro Diario) de Juan de Lerma.	1541-51	2	journal	YES	LERMA, Juan de	127	447/295
2	Manual del libro de compras de Diego de Bernuy.	1546-55	4	journal	NO	BERNUY, Diego de	314	443x300
3	Manual (o Libro Diario) de Diego de Bernuy.	1546-50	5	journal	YES	BERNUY, Diego de	182	443x290
4	Libro de Caja [o Mayor] de García de Salamanca.	1548-52	7 ledger		NO	SALAMANCA, García de	206	467/300
5	Manual (o Diario) de Miguel y García de Salamanca.	1548-51	6 journal		NO	SALAMANCA, Miguel y García de	107	467/300
6	Libro de Caja [o Mayor] de García de Salamanca.	1552-63	9 ledger		YES	SALAMANCA, García de	298	467/295
7	Manual (o Diario) de García de Salamanca	1552-63	10 journal		NO	SALAMANCA, García de	not available	not available
8	Libro grande de Caja (o Mayor) de García y Miguel de Salamanca.	1564-69	11 ledger		NO	SALAMANCA, Miguel y García de	115	457/300
9	Libro de caja (o Mayor) de Quintanabarbos.	1571-78	17 ledger		NO	QUINTADUENAS, Pedro de	207	457/295
10	Manual del libro grande de Caja de García y Miguel de Salamanca.	1564-69	16 journal		NO	SALAMANCA, Miguel y García de	61	467/300
11	Manual del libro de Nota de García de Salamanca.	1545-57	20 journal		YES	SALAMANCA, García de	168	337/210
12	Manual del libro en el que se asientan las mercadurías de Ruan.	1551-58	22 journal		YES		190	337/220
13	Libro de Notas de García de Salamanca.	1553-63	24 ledger 1		YES	SALAMANCA, García de	220	297/210
14	Manual (o Diario) del libro de Notas y cosa asuntos de García de Salamanca.	1554-63	25 journal		YES	SALAMANCA, García de	122	337/210
15	Manual de las cuentas que tenía con varias personas la compañía de García y Miguel de Salamanca.	1555-63	27 ledger		YES	SALAMANCA, Miguel y García de	36	337/210
16	Manual (o Diario) del libro de los negocios de la compañía de García y Miguel de Salamanca.	1560-61	29 journal		YES	SALAMANCA, Miguel y García de	130	337/230
17	Manual del libro de cierre de Miguel y García de Salamanca.	1561-64	31 journal		YES	SALAMANCA, Miguel y García de	22	337/210
18	Manual (o Diario) de los negocios de la compañía de García y Miguel de Salamanca.	1563-63	32 journal		YES	SALAMANCA, Miguel y García de	130	337/240
19	Libro de cierre de la compañía de los herederos de Bernardino Vallejo y Ventura del Castillo.	1573-78	33 ledger		YES	VALLEJO, Bernardino de y CASTILLO, Ventura del	191	317/215
20	Libro de los negocios de García y Miguel de Salamanca	1562-63	34 not specified		YES	SALAMANCA, Miguel y García de	190	335/240
21	Manual del libro de los negocios de compañía de García y Miguel de Salamanca.	1566-74	36 journal		YES	SALAMANCA, Miguel y García de	117	347/240
22	Libro en el que se asientan las mercaduras de Ruan.	1551-57	36 not specified		NO		not available	337/220
23	Manual del libro de Alonso de Arlanzón.	1572-78	45 journal		YES	ARLANZÓN, Alonso de	106	347/230
24	Libro de ningos, segunia y otras cuentas particulares de Alonso de Arlanzón.	1573-78	47 not specific		NO	ARLANZÓN, Alonso de	427	347/230
25	Manual del libro de cierre de los herederos de Bernardino de Vallejo y de Ventura del Castillo.	1573-79	49 journal		YES	VALLEJO, Bernardino de y CASTILLO, Ventura del	106	317/215
26	Manual (o Diario) de Ningos de García de Salamanca.	1565-66	64 journal		YES	SALAMANCA, García de	40	337/210
27	Libro de ningos de Luis de Salamanca.	1565-66	70 ledger		YES	SALAMANCA, Luis de	62	337/210
28	Manual del libro de Ningos de Alonso de Arlanzón.	1573-78	71 journal		YES	ARLANZÓN, Alonso de	226	347/235
29	Libro de Caja de Alonso de Arlanzón.	1573-78	100 ledger		partially photographed	ARLANZÓN, Alonso de	192	347/233
30	Libro de los negocios de compañía de García y Miguel de Salamanca.	1565-65	106 ledger		partially photographed	SALAMANCA, Miguel y García de	201	347/230
31	Libro de cierre que García y Miguel de Salamanca tenían con varios sujetos.	1561-62	109 ledger		NO	SALAMANCA, Miguel y García de	42	337/235
32	Libro de cuenta y razón de los negocios que tenía la compañía de García y Miguel de Salamanca, así de ningos como de mercaduras.	1560-61	110 ledger		partially photographed	SALAMANCA, Miguel y García de	206	337/230
33	Libro de Caja [o Mayor] de los negocios de García y Miguel de Salamanca.	1562-65	111 ledger		NO	SALAMANCA, Miguel y García de	118	337/230
34	Libro de los negocios que hacen en compañía García y Miguel de Salamanca.	1560-74	789 ledger		NO	SALAMANCA, Miguel y García de	185	347/235

Table 1:
List of Accounting Books kept in ADPB

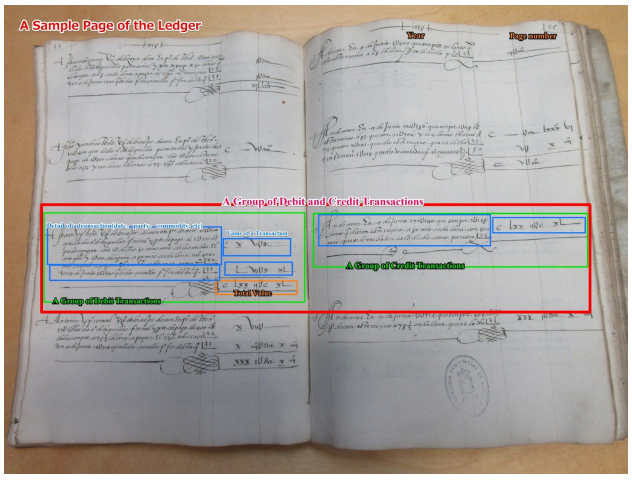


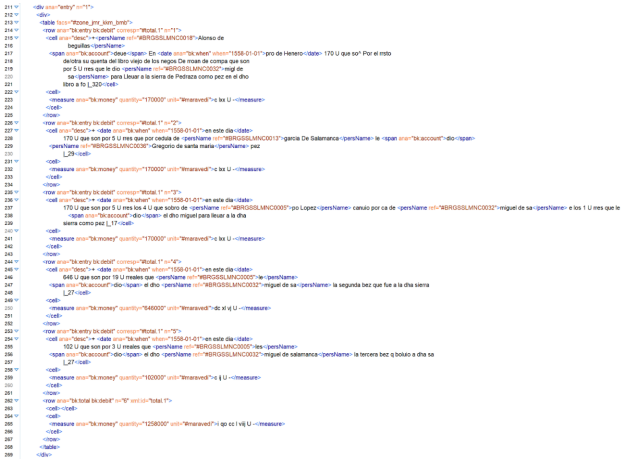
Figure 1:
Credit and debit entries in the ledger

Analytical Framework

For the markup of accounting records, XBRL might be a possible candidate framework (Anderson et al. 2016). However, this modern framework seems to be not suitable enough to capture all the elements of the historical records, not present in the contemporary ones. Instead, we have chosen DEPCHA, an extended TEI schema. DEPCHA has developed a domain ontology, based on the concept of 'Transactiongraphy' (Tomasek and Bauman, 2013), for extracting machine-readable data in RDF format from transactional information in accounting records in the TEI or CSV format. A member of the DEPCHA, states that their future challenge is to normalize the vocabulary of words such as "currency" and "item" in accounting records that emerge from various case studies and to build an interoperable data model for open data (Pollin, 2019), but at present, there are not so many case studies. Furthermore, concrete examples of double-entry books, which are considered as the basis of accounting practice in the modern world, have not sufficiently been marked up using the DEPCHA scheme. Our research will contribute to filling this gap and giving practical feedback to DEPCHA. We also think that more normalized TEI/DEPCHA can make the analysis of historical financial data more efficient, when the scale of data becomes bigger.

Progress and prospect

Let us explain our markup policy. As mentioned earlier in the materials section, the correspondence between transactions is displayed across each of the left and right pages, so the facing page needs to be structured as a coherent space when marking up. (cf. Markup 1). While following DEPCHA's recommended policy of using <table> tags to structure the transaction information in a tabular form, we ensure that the @ana attribute value is entered in a human-readable form—though it is not completely machine-processable at the moment—, so that the correspondence between the left and right pages is clearly indicated. In addition, the reference relationships between the diary and ledger are also structured (cf. Markup 1).



Markup 1:
Tagging an Entry in the Ledger

We will demonstrate our achievements using the above-mentioned mark-up files to conduct analyses, as follows: (1) identification of abbreviated person names, (2) accounting terms and abbreviations, and (3) checking the balance of payments; all of them is available to browse at our page .

We prospect to mark up more accounting books kept in the same archive, and launch trend and network analyses using metadata, e.g. time, name of person/company, place, commodity, and price. And we will also investigate more proper way of displaying our analysis. We would like to share our framework with the researchers pursuing similar themes to improve the analytical frameworks and tools for the study of economic history.

Bibliography

Anderson, C., Eide, O., Orlowska, A., Pindl, K., Tomasek, K. and Vogeler, G. (2016). Modeling semantically Enhanced Digital Edition of Accounts (MEDEA) for Discovery and Comparison on the Semantic Web. Humanities Commons. https://hcommons.org/deposits/item/hc:24347/ (accessed 21 April 2022).

GAMS (2017). DEPCHA - Digital Edition Publishing Cooperative for Historical Accounts *DEPCHA - Digital Edition Publishing Cooperative for Historical Accounts* http://gams.uni-graz.at/archive/objects/context:depcha/methods/sdef:Context/get?mode=howto (accessed 1 June 2019).

Pollin, C. (2019). Digital Edition Publishing Cooperative for Historical Accounts and the Bookkeeping Ontology. *Proceedings of the Doctoral Symposium on Research on Online Databases in History*, http://ceur-ws.org/Vol-2532/paper1.pdf (accessed 21 April 2022).

Tomasek, K. and Bauman, S. (2013). Encoding Financial Records for Historical Research. *Journal of the Text Encoding Initiative* (Issue 6), http://journals.openedition.org/jtei/895 (accessed 21 April 2022).

Sound Iconicity and Digital Humanities. A Case Study of Spanish Golden Age Theatre

Kroll, Simon
simon.kroll@univie.ac.at
University of Vienna, Austria

Recently, the relevance of sound for the meaning of words in a language has increasingly gotten scholarly attention. Works on sonic iconicity in different languages all over the world are gaining importance, regarding both the importance of sound in bound language (lyrics), and in daily spoken language. This research has provided serious evidence suggesting that the sound of a word already gives us information about its meaning, or at least, about the overall associations attached to it. In other words, phonemes already seem to carry semantic information (Auracher, 2020). Whereas a number of studies have investigated phonosemantics concerning single phonemes in a general sense, a noticeable research gap exists in the context of poetry or verse drama. Since it became the first modern mass culture, the early modern verse drama appears to constitute an especially crucial corpus for this field of research.

The short presentation format was chosen to present the ongoing work of my research including first hints on the use of phonosemantic relations in Spanish early modern theatre (16 th and 17 th century). As a basis for researching such relations, in a first step, we created a corpus from the existing digital text repositories, collecting over 500 plays by different playwrights such as Pedro Calderón de la Barca, Lope de Vega, Tirso de Molina, Sor Juana Inés de la Cruz, Mira de Amescua, and many others. The second step included the development of a Python script, enabling the analysis of the phonic structure of every single verse line. The program identifies the number of syllables, the rhythmical patterns, the rhymes, identifies between stressed and unstressed vowels, and creates a csv-file to collect all the created data. As a first result, a phonologic transcriptor and a syllabic analysis tool have already been published as Python libraries (see Sanz Lázaro, *fonemas* and *silabeador*). These scripts will be presented briefly, taking into account the existing research on automatic verse analysis in Spanish (González-Blanco, Remón, de la Rosa).

In order to analyze these phonetic data on a large scale, a number of shorter Python scripts were developed. Using the libraries pandas and matplotlib, the goal of research is to answer the following questions: Do the different playwrights have preferences for different rhythmical patterns? Do these rhythmical patterns appear randomly throughout the texts? Or is it possible to establish a pattern between negative and positive emotions and the occurrence of certain rhythmical structures? To what extent are rhythmical patterns related to other phonosemantic phenomena? Is there an interrelation between rhyming structures and different rhythms or a connection between the stressed vowels and the rhythmical patterns?

As it aims to evoke very strong emotions in its spectators, the early modern theatre is often referred to as an affect machinery. This aspect of Spanish Golden Age theatre is usually reduced to visual aspects: the use of baroque theatre machinery, special effects, and the overwhelming costumes of the professional actors.

In contrast, this research shows, that an important medium to achieve this goal can be found in the sound of the verses themselves. Therefore, this presentation has a threefold aim:

- to present the new Python scripts and libraries created for the automatic verse analysis;
- to discuss phonosemantic relations and the digital methods to analyze them;
- to provide new insights into the importance of sound effects in the affect machinery of the early modern theatre.

Thus, the presentation will show important advances in the automatic analysis of metrical texts, which can easily be transferred to texts from other epochs and, with slight adaptations, also to other romance languages like Italian.

Bibliography

Auracher, J., Menninghaus, W. and Scharinger, M. (2020). Sound Predicts Meaning: Cross-Modal Associations Between Formant Frequency and Emotional Tone in Stanzas. *Cognitive Science*, **44**:10: e12906. https://doi.org/10.1111/cogs.12906.

Bird, S., Klein, E. and Loper, E. (2009). *Natural Language Processing with Python*. Sebastopol, CA: O'Really.

De la Rosa, J., Pérez, Á., Hernández, L., Ros, S. and González-Blanco, E. (2020). Rantanplan, fast and accurate syllabification and scansion of spanish poetry. *Procesamiento del Lenguaje Natural* **65**, 83-90.

Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R: A Package for Computational Text Analysis. *The R Journal* **8,1**: 107–21.

González-Blanco, E. Pérez, Á., Ros, S. *PoetryLab. An Open Source Toolkit for the Analysis of Spanish Poetry Corpora.*

Karsdorp, F., Kestemont, M. and Riddell, A. (2021). *Humanities Data Analysis: Case Studies with Python*. Princeton, Oxford: Princeton University Press. https://press.princeton.edu/books/hardcover/9780691172361/humanities-data-analysis.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60. Baltimore, Maryland: Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-5010.

Marco, G. De La Rosa, J., Gonzalo, J. Ros, S., González-Blanco, E. (2021). Automated Metric Analysis of Spanish Poetry: Two Complementary Approaches. *IEEE Access* **9**, 51734-51746.

McKinney, W. *Python for Data Analysis*, Sebastopol, CA: O'Really.

Plecháč, P. (2018). Versification and authorship attribution. A pilot study on Czech, German, Spanish and English poetry. *Studia Metrica et Poetica*, **5, 2**: 30-54.

Remón, GM., Gonzalo, J. (2021). Escansión automática de poesía española sin silabación. *Procesamiento del Lenguaje Natural* **66**: 77-87.

Sanz Lázaro, F. (2021). *Fonemas. A Python phonologic transcription library for Spanish*. Version 1.0.3. fsanzl, Okt. 2021. Software. *GitHub*, https://github.com/fsanzl/fonemas

Sanz Lázaro, F. (2021). *Silabeador: A Python library for syllabic division and stress detection for Spanish*. Version 1.0.5, fsanzl, Okt. 2021. Software. *GitHub*, https://github.com/fsanzl/silabeador.

Toward an Affordances Approach to Literacy in the Digital Humanities

Kulkarni, Kavita
kavitak@princeton.edu
Princeton University, United States of America

Toward an Affordances Approach to Literacy in the Digital Humanities
This paper reflects on the author's experience designing and teaching *Humanistic Approaches to Media and Data*, an inaugural digital-humanities course included in the

STEM course offerings in the Freshman Scholars Institute at Princeton University in the summer of 2021. The course was designed to teach critical media and data literacy through a hybrid structure of reading seminar/programming laboratory. This paper makes a case for an "affordances" approach to literacy-oriented pedagogy in the digital humanities, arguing that the ability to "read"—by "working through"—the affordances of DH tools and techniques marks a logical move in the progression within the critical literacy paradigm from textual literacy to media literacy to data literacy. [1] This paper focuses in particular on the progression from media literacy to data literacy.

The paper starts with the premise that practices of meaning-making are both the object of study and mode of knowledge production in the humanities, and that these practices have become increasingly complex over time by way of technological advancements in communication media. Likewise, the contours of literacy and of pedagogy centered on literacy have shifted with regard to their focus and objective. The flourishing of "media literacy" pedagogy in the 1990s, for example, offered an addendum to traditional notions and pedagogies of literacy that focused on the reading and writing of language. Media literacy added to these skills the understanding of how media systems operate for the purpose of safeguarding against mass media manipulation and producing discerning news consumers. This approach gained heightened relevance in the United States during and after the 2016 presidential election, when "fake news" became a popular framework for addressing the phenomenon of misinformation being circulated in the public sphere. In a report published by Data & Society in 2018, the authors note that in this new era of intensified misinformation, "[m]edia literacy has become a center of gravity for countering 'fake news,' and a diverse array of stakeholders—from educators to legislators, philanthropists to technologists—have pushed significant resources toward media literacy programs" (Bulger and Davison, 2018). Incidentally, a blog post published a month later by Data & Society's very own danah boyd warned, "[i]f we're not careful, 'media literacy' and 'critical thinking' will simply be deployed as an assertion of authority over epistemology" (boyd, 2018). In other words, "when youth are encouraged to be critical of the news media, they come away thinking that the media is lying" (boyd, 2018).

To offer a way around this standoff in media literacy discourse regarding what it means to teach critical thinking in the age of "fake news," and to steer the conversation toward a consideration of what critical literacy looks like in the age of ubiquitous data and computation, this paper proposes an "affordances" approach to literacy, building on existing literature in DH pedagogy that argues for programming/making/building as a form of learning. In particular, this "affordances" approach to literacy is premised on three beliefs: 1) It is valuable for students to start with their own research questions and explore how DH can or cannot assist them in exploring these questions, instead of starting with DH tools or datasets for which they would have to find an application; 2) It is important to make room for failure, as it is a prerequisite to students understanding technological design as a matter of affordances and limitations; and 3) To ensure that literacy is in fact "critical," it is important to draw connections between affordances and social or political values, emphasizing how design in media and computational technologies is not only a matter of function, but also principles like accessibility, privacy, and collectivism.

Bibliography

boyd, d. (2018). You think you want media literacy… do you? *Data & Society: Points*, https://points.datasociety.net/you-think-you-want-media-literacy-do-you-7cad6af18ec2 (accessed 11 December 2021).

Bulger, M. and Davison, P. (2018). The promises, challenges, and futures of media literacy. *Journal of Media Literacy Education*, **10** (1): 1–21.

Gibson, J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton, Mifflin and Company.

Notes

1. This paper draws from the work of James J. Gibson—long utilized in cognitive psychology, design studies, and human-computer interaction—for its definition of affordance, which for the purposes of this abstract will be abridged to "a value-rich ecological object."

Visualizing Academic Networks and Trends through Acknowledgements: Japanese Scholars in Islam-related Studies

Kumakura, Wakako
kumakurawakako@gmail.com
ILCAA, Tokyo University of Foreign Studies

Sunaga, Emiko
sunaga.emiko@gmail.com
U-PARL, The University of Tokyo

The purpose of this study is to examine the validity of acknowledgement networks in reconstructing academic networks and academic trends. As for the acknowledgement network, its usefulness to draw "communication networks" between researchers was suggested in a study using acknowledgements included in journal articles (Mushanokoji, 1982). Since that time, attention has been paid from various disciplines to the meaning of the act of sending acknowledgements and the relationships extracted from it, yet studies that visualize these relationships as a network map and examine their validity are still inadequate. We ran experiments reconstructing academic networks based on the acknowledgements included in doctoral theses. In the humanities community in Japan, there is still a strong tendency for the publication of a single-authored work to be evaluated as an achievement rather than a co-authored article. Under such circumstances, it is significant to demonstrate the validity of the acknowledgement network as an alternative method of visualizing academic networks to the co-authorship network.

The source materials used for this study are 120 Islam-related doctoral theses submitted to Japanese universities from the 1950s through the 2010s, as recorded in the CiNii Dissertations database (https://ci.nii.ac.jp/d/). The number of Islam-related doctoral theses in Japan began to rise in the 1980s and continued climbing into the 2010s (Fig. 1). Behind this increase in Islam-related doctoral theses seen from the 1990s onward was the introduction of an innovative research platform that transcended the boundaries of the university. In the 2000s, five leading Japanese universities and institutions launched a large-scale collaborative research project with many young researchers participating in both the research and administrative aspects. Moreover, university-level Islam-related studies —previously centered on history courses—shifted to inter-disciplinary studies under this project's "Islamic Area Studies" framework, a phenomenon reflected by increases in area studies and multi-disciplinary research (Miura, 2004).

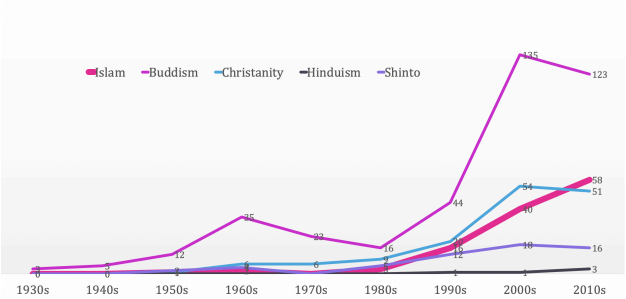


Fig. 1:
Number of doctoral theses in Japan (by religion studied)

To visualize these social network connections, a graph is created displaying an "acknowledgement network," placing authors and the scholars acknowledged by those authors as nodes, with edges extending from the acknowledgee to the author (Tian et al., 2021). The area of specialization for each person was taken from records in the KAKEN database of Japanese researchers (https://nrid.nii.ac.jp/index/). This makes it possible to visualize trends for separate academic fields.

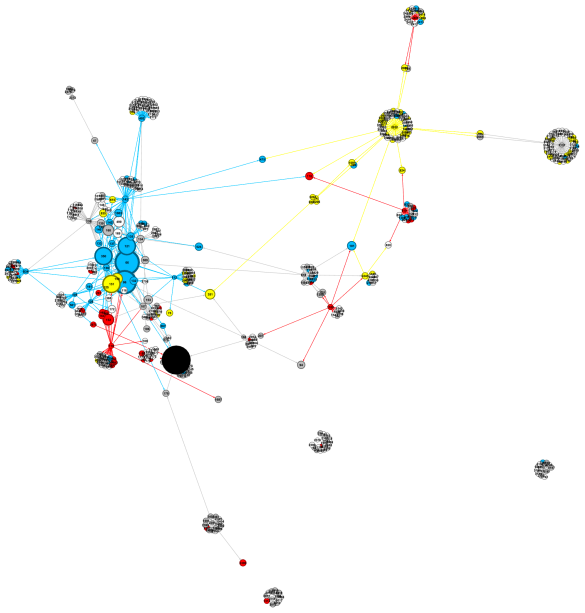


Fig. 2:
Acknowledgement network since the 2010s

Figure 2 shows the resulting acknowledgement network, created using Gephi network visualization software (https://gephi.org/). The number in each node is ID. This visualization of the acknowledgement network based on the doctoral theses since 2010 shows the presence of area studies (in blue) in addition to the history studies (in red) that were considered to be the mainstay of academic circles. Some acknowledgees are connected to multiple authors, indicating that these researchers are considered influential in their academic circle. These acknowledgees comprise the generation of researchers that contributed to the growth of Japan's "Islamic Area Studies" network in the 2000s and have some formal or informal teacher-student connection to the authors. From the above, we can say there is a definite validity in using an acknowledgement network to visualize academic networks and the steps in establishing and developing academic connectivities. Also, the acknowledgements in the theses, which previously cited mainly the author's family and academic advisors, have

since the 2010s grown to list the names of as many as 10 or more researchers on average. These quantifiable trends reflect not only the increased opportunities for academic interactions as well as more researchers contributing to the completion of doctoral theses but also the given author's intention to develop ever more public associations with other researchers. This is likely indicative of their aim to enhance their academic reputation by connecting with a larger number of prominent researchers.

Bibliography

武者小路, N. (Mushanokōji, N.) (1982). 「アメリカ経済史研究者間のコミュニケーション・ネットワーク：謝辞による分析」(Communication Networks among Scholars of American Economic History: An Analysis through Acknowledgments)『図書館学会年報』(*Annals of Japan Society of Library Science*), 28 (1): 43–45.

Miura, T. (2004). "Survey of Middle East Studies in Japan: Historical Development, Present State, and Prospectus," *Japan Association for Middle East Studies*, 19 (2): 169–200.

Tian, S., Xu, X. and Li, P. (2021). "Acknowledgement network and citation count: the moderating role of collaboration network," *Scientometrics*, 126: 7837–7857, https://doi.org/10.1007/s11192-021-04090-y (accessed 19 April 2022).

Diary of our initiatory journey on the continent of data citation in SSH

Larrousse, Nicolas
nicolas.larrousse@huma-num.fr
Huma-Num, CNRS, France

Gray, Edward
edward.gray@dariah.eu
Huma-Num, CNRS, France

Concordia, Cesare
cesare.concordia@isti.cnr.it
ISTI, CNR, Italy

Diary of our initiatory journey on the continent of data citation in SSH

The metaphor of a travel journal of an expedition seemed appropriate to us to present this work carried out during the SSHOC [1] project.

The first part was to study this terra incognita by making an inventory of citation practices [2]. To summarize, we discovered that in the SSH research communities we investigated, practices were seldom standardized and were very diverse, generally producing citations that could not be processed by machines: in other words they were not "actionable".

This led us to develop a sort of guide necessary to journey through this new, uncharted territory in the form of a set of recommendations [3] to build citations in SSH. So as not to reinvent the wheel, we based these recommendations on existing principles created by Force11 [4] by adapting them to the specific characteristics of the SSH data. These recommendations were validated by a committee of experts from different backgrounds and structures (RDA participants, CODATA director, OpenAire Engineers etc.) during a round table [5] and in a parallel review process.

Then we decided to analyze the resources available in this new territory, that is, the repositories that are so crucial to be able to cite data. We carried out an analysis of 85 repositories against 7 quality criteria to address the "challenges" described in the recommendations mentioned above:

- **PID** from "Unique Identification & Persistence"
- **Landing page** from "Access"
- **Structured metadata** from "Importance & Credit and Attribution"
- **Cite as** from "Evidence, Specificity & Verifiability"
- **Versioning** from "Specificity and Verifiability"
- **Standardized vocabularies** from "Interoperability and Flexibility"
- **Links to publications** from "Importance"

The results of this survey [6] are encouraging - even if there is room for improvement, particularly in the use of Persistent Identifiers. Importantly, the presence of a landing page in almost all cases allowed us to build up a test sample made up of a very diverse dataset from those repositories for which we want to build standardized and actionable citations.

In parallel we developed a tool in order to "harvest" the resources found in this new land so as to better understand them and also be able to explain them to others. We developed a prototype composed of three components:

- a harvester which grabs information about a dataset and normalizes it based on the work done by SCHOLIX [7]
- an API to disseminate the metadata of the citation thereby making it actionable
- a citation viewer for human purposes

For the first iteration to populate this prototype, we used the dataset collected during our survey of repositories and we are going to gradually add more datasets from various sources.

This prototype is primarily designed to implement what we called “actionability” to a citation and provide a ready-to-use citation in various citation formats. Starting from the PID of a dataset, the prototype attempts to aggregate metadata from different sources: the repository of the dataset, the PID Registration Agency and a number of Knowledge Graphs. For instance, while metadata associated with a DOI (Digital Object Identifier) are limited and those provided by a handle are even more scarce, it is possible to get more information from a landing page and thus enrich the citation.

We also used another indirect approach to gather additional information by using a registry of repositories (RE3Data [8]) which provides, among other things, information on the available APIs available for a specific repository.

Thus, the prototype can give a unified view of information about datasets coming from different sources. For researchers, it thus avoids cumbersome work on how to cite a dataset or get information about its provenance. In return, it makes a researcher aware of the importance of properly documenting a dataset and depositing it in a “good” repository.
The code of the prototype is available on the GitLab instance maintained by ISTI-CNR.

This paper will present in greater detail what we learned at **each step of this expedition** and how a research project can take advantage of a good citation system to enhance the visibility of the output. We will also introduce the potential uses based on the information provided by the prototype such as the possibility of associating a specific tool to process data or the use of this information as a base to build data papers.

Bibliography

Blaney, Jonathan. (2012). 'The Problem of Citation in the Digital Humanities'. In: Clare Mills, Michael Pidd and Esther Ward. *Proceedings of the Digital Humanities Congress 2012*. Studies in the Digital Humanities. Sheffield: The Digital Humanities Institute, 2014. Available online at: https://www.dhi.ac.uk/openbook/chapter/dhc2012-blaney

Rauber, Andreas, Asmi, Ari, van Uytvanck, Dieter, & Proell, Stefan. (2015). Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC). https://doi.org/10.15497/RDA00016

Task Group on Data Citation Standards and Practices, C.-I. (2013). Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. *Data Science Journal*, 12, pp.CIDCR1–CIDCR7. DOI: http://doi.org/10.2481/dsj.OSOM13-043

Notes

1. https://sshopencloud.eu/
2. https://doi.org/10.5281/zenodo.3595965
3. https://doi.org/10.5281/zenodo.5361717
4. https://doi.org/10.25490/a97f-egyk
5. https://www.sshopencloud.eu/news/roundtable-experts-data-citation
6. https://doi.org/10.5281/zenodo.5603306
7. http://www.scholix.org/
8. https://www.re3data.org/

A unified path from data to publications: Three French infrastructures in the COMMONS project

Larrousse, Nicolas
nicolas.larrousse@huma-num.fr
Huma-Num, CNRS, France

Pellen, Marie
marie.pellen@openedition.org
OpenEdition, CNRS, France

Roux, Dominique
dominique.roux@unicaen.fr
Metopes, Pôle Document Numérique, CNRS & Université de Caen, France

Baude, Olivier
olivier.baude@huma-num.fr
Huma-Num, CNRS, France

In order to support research projects in SSH (Social Sciences and Humanities), France has developed centralized infrastructures which are listed on the French national roadmap for infrastructures: this means that they are financed over the long term.

Three of these infrastructures are especially complementary:

Huma-Num (See https://www.huma-num.fr and https://documentation.huma-num.fr/humanum-en/) provides a set of platforms, tools and support for the

processing, conservation, dissemination and long-term preservation of digital research. OpenEdition (See https://www.openedition.org) provides a set of four scholarly communication platforms and Metopes (See https://www.metopes.fr) provides a set of tools and methods, built according to single source publishing model, enabling the creation of natively structured editorialized contents.

Even though these infrastructures provide all the services required throughout the life cycle of research, their different environments are separate, making it necessary to switch from one platform to another, for instance to carry out the process of linking data and publications.

In the context of the increasing development of Open Science policies, both at national and European levels, these three main infrastructures joined their forces to build a project to address these challenges. The main focus of the COMMONS [1] project (COnsortium of Mutualised Means for OpeN data & Services for SSH) is of course to develop bridges between the three infrastructures articulated around the link between data and publications but it also covers various other related aspects.

A key step to achieve the project goals is to develop communication modules between the different platforms which represent the foundation on which the other achievements will be based. For instance, one of the most important components to build this environment is to use a common authentication system which allows users to connect to any platform with the same credentials. This requires technical adaptations but also substantial preliminary work to ensure coherence between our lists of respective users.

Similarly, to link data and publications, it will be necessary to establish bi-directional links between the data repository provided by Huma-Num and the scholarly communication platforms from OpenEdition associated with tools from Metopes. Interoperability modules will be developed to implement these links associated with a specific workflow aiming to maintain consistency of information between the different platforms. For example, one of the key challenges will be to take into account the synchronization aspects when establishing the link between the data and the publication.

But of course, it is not sufficient to implement these technical requirements. It is also necessary to accompany the various categories of users in the use of this new system. Therefore, an important part of the project is to model and implement a seamless user experience associated with extensive training sessions adapted to all the stages of the research lifecycle.

Alongside these three major axes which structure the project, there will be a strong focus on data, particularly what are called “sensitive data” which are frequently used for research in SSH. In this case, the requirements between the need for strong security on the one hand, and the need to provide research projects with easy access to data on the other hand, are contradictory: the possibility of providing access to protected or sensitive resources is one of the important issues of the project. Finally, as connection and interoperability with other infrastructures at national and international levels is a key part of the project, APIs respecting general standards will be developed to be able to disseminate at large the information generated by the project.

The main expected outcome, besides making the publication process more fluid for research projects, is to foster the visibility of French research. This will be achieved mainly by improving the quality of its contents by following the FAIR [2] principles and the development of research project skills. Another outcome is to have available all the bricks necessary for the construction of new types of complex publications in the context of rapidly evolving new forms of scientific writings, in particular data papers or actionable papers.

This paper will develop both the technical and organizational aspects of this ambitious project, which will require highly specialized professional profiles to support both producers (researchers, engineers, editors) and users (researchers, students, companies, public authorities, media) of content, and will last 8 years for a total cost around 30 M €.

Bibliography

Colavizza G, Hrynaszkiewicz I, Staden I, Whitaker K, McGillivray B (2020) The citation advantage of linking publications to research data. PLoS ONE 15(4): e0230416. https://doi.org/10.1371/journal.pone.0230416

Burton, Adrian, & Koers, Hylke. (2016). ICSU-WDS & RDA Publishing Data Services WG Interoperability Framework Recommendations (1.0). https://doi.org/10.15497/RDA00002

Notes

1. See https://www.cnrs.fr/en/node/5607
2. https://www.go-fair.org/fair-principles/

Electronic Literature: thinking a taxonomy

Lescouet, Emmanuelle
emmanuelle.lescouet@umontreal.ca
Université de Montréal, Canada

Vitali-Rosati, Marcello
marcello.vitali.rosati@umontreal.ca
Université de Montréal, Canada

Documenting the digital literature represents a research challenge (Vitali-Rosati, 2017) because it requests the institution of modularity in digital taxonomies, on a field in constant evolution (Hayles, 2007). The fast evolution of media and their global diffusion give access to many technologies and tools. The latter brings vast diversity as much on the practices of reading as on physical manipulations: the digital literature covers at the same time network literature, interactive fictions, digital books, virtual reality experiments or notifictions (Bouchardon, 2009 and 2012; Lescouet, 2021).

The very existence and dissemination of literature build a trans-platform, inclusive and plastic culture (Ryan, 2015). It challenges researchers to find and identify artworks. These reading's materialities are evolving beyond the fictional 'magic circle' often took in consideration in reading theories (Huizinga 1938; Picard 1986; Mace, 2011). Hyperconnection overlaps with contemporary tools and creates geographical and geolocalized fragmentation, as well as temporal fragmentation. This allows a unique user immersion, that is up to us to document in its current and permanent developments.

These creations find their audience in particular online communities. For this reason, they could be difficult to find and to present to a larger public, academic or not. However, the institutionalization of these artworks and their qualification allow us to study them: they become analyzable because they belong to an academic environment. Therefore, we must provide new ways to talk about them. It is imperative to establish a common vocabulary (see the early 2000s CELL Project research on this issue), a common language, to unify an approach and a documentation between the existing databases. It is at the same time a necessity and an obstacle in the conservation of the particularities of each project.

Nonetheless, we observe common trends as well as common descriptors. This recognition has led us to build a thesaurus within the inter-university partnership *Littérature Québécoise Mobile*. Starting from there, we have then established collaborative vocabularies in OpenTheso. The urgency of working together to establish definitions and hierarchies of vocabularies requires the institution of equitable governance and a tool to allow discussions and linkage between artworks.

Building a thesaurus was therefore imperative. We must build an ontology in the technical sense of the term, which was structured and shareable. The descriptors had to be included, structured, editable and manipulable, but above all they had to be linkable to the descriptive files of the artworks studied. The choice of SKOS appeared obvious. Moreover, it was necessary to choose a tool that would allow to manage and store the permanent id of the terms to stabilize the descriptors. Opentheso allows this. Created in 2005 by the Federation and Resources on Antiquity, it is now part of the tools hosted by Huma-Num.

This proposal, without being exhaustive, focuses on three main aspects of this documentation:

The technical aspects and reading supports of the artworks in the corpus; The possible interactions between the reader and the artwork established through the gestures of readings or through the experimentation implemented, in duo with the principles of the artwork's internal organization; The genres and literary forms convened, in the diversity and the flexibility that such a position implies.

The aim of this documentation for digital artworks is to understand the close interaction between them and their influence on the reception of the artworks. The influence of the techniques concerning the presence and the incorporation of the body-reader in the propositions allow us to study the immersion and the temporalized lectorial propositions or the disciplinary borders between the different digital artistic disciplines, notably between video game and literature.

To do so, we will study in detail how such taxonomies could be set up and how the collaborative work around them is organized. We will then confront the studied fields with the real corpus (on the questions of 'corpus place', see Emerit, 2016) in order to consider the limitations and constraints encountered.

Bibliography

Bouchardon, S. (2009) *Littérature numérique : le récit interactif*. Paris: Hermès Science.

Bouchardon, S. (2012) 'Du récit hypertextuel au récit interactif', *Revue de la BNF*, (42), pp. 13–20.

Collectif (no date) *CELL Project*. Available at: https://cellproject.net/ (Accessed: 19 October 2020).

Emerit, L. (2016) 'La notion de lieu de corpus : un nouvel outil pour l'étude des terrains numériques en linguistique', *Corela*, 14(1). doi:https://doi.org/10.4000/corela.4594.

Hayles, K.N. (2007) *Electronic literature: what is it?* Available at: https://eliterature.org/pad/elp.html (Accessed: 2 April 2019).

Huizinga, J. (1938) *Homo ludens. Essai sur la fonction sociale du jeu*. Translated by C. Sérésia. Paris: Gallimard (Les Essais).

Lescouet, E. (2021) 'La notifiction : ancrer le livre dans le quotidien', in *Études du livre au XXIe siècle*. Québec, Canada. Available at: https://projets.ex-situ.info/

etudesdulivre21/liv5/lescouet/ (Accessed: 15 December 2021).

Macé, M. (2011) *Façons de lire, manières d'être*. Paris: Gallimard (NRF essais).

Picard, M. (1986) *La lecture comme jeu : essai sur la littérature*. Paris: Les Éditions de Minuit (Critique).

Ryan, M.-L. (2015) *Narrative as Virtual Reality 2*. Baltimore: Johns Hopkins UP.

Vitali-Rosati, M. (2018) *La littérature numérique francophone : enjeux théoriques et pratiques pour l'identification d'un corpus*, *Culture numérique. Pour une philosophie du numérique*. Available at: http://blog.sens-public.org/marcellovitalirosati/la-litterature-numerique-francophone-enjeux-theoriques-et-pratiques-pour-lidentification-dun-corpus/ (Accessed: 22 November 2021).

Digitizing and Recovering the Knowledge of Traditional Chinese Colour of the Nanjing Brocade

Lin, Wensi
vencylin@qq.com
Shanghai Museum, China

Chen, Jing
cjchen@nju.edu.cn
Nanjing University, China

Zhang, Mengyue
zhangmengyue_boao@qq.com
Nanjing Boao Culture Technology Company Limited, China

Wang, Jisheng
513923540@qq.com
The Institute of Archaeology, Chinese Academy of Social Sciences, China

Li, Mengqi
manchi.li@traditionow.com
Shanghai Traditionow Culture Development Corporation Limited, China

The Chinese handicraft, Nanjing Brocade, is a traditional luxury silk served for royal and noble family during the 13th to the 20th century. It was weaved from polychrome yarns and represented the topmost craftsmanship of natural dyeing and weaving in China (Huang et al., 2003). The ancients created hundreds of colours by over-dyeing, and used numerous terms to describe the varied colours, reflecting their perception and congnition towards colours. However, colour knowledge of Nanjing Brocade has not been systematically and precisely documented. Colour cognition or dyeing technique was mainly kept as tacit experience knowledge and passed on from master to apprentice privately. Hence, Chinese colour knowledge has been forgotten slowly. And with the ancient fabrics irreversibly fading, it's difficult to tell what colours they once were, and how to reoccur them. Nowadays, there are projects trying to recover the ancient textiles colour, but most of the time remained in traditional way of literature interpretation and experience induction (Kim, 2006; Liu, et al., 2020). A comprehensive study using digital methodology on Nanjing Brocade colour is in lack. We can learn from researchers in other fields, such as Li (2019) who explored the colour of ancient Chinese buildings basing on the analyses of literature, cultural relics, and techniques, with the experiments of pigments remaking.

This research introduces a comprehensive research on the colour knowledge of Nanjing Brocade that leverages information from literature, artifacts, and expert experience. Based on the digital methodology of text mining, dyeing experiment, and perceptual evaluation, the research explores a way of knowledge production that produces systematic, accurate, and standard scientific knowledge. The research results demonstrate the colour range of Nanjing brocade from both conceptual and visual perspectives. The main process is shown in Fig. 1.

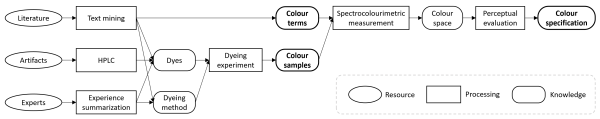


Fig. 1:
The knowledge production process of Nanjing brocade colour

Firstly, in order to solve the problem that traditional literature knowledge is scattered, we created a database for the relevant literature and collected information of colour terms through text mining (Tan, 1999). MARKUS, a text analysis and reading platform, was used to label the colour terms of Nanjing brocade and other entities such as its pattern, date and wearer's identity. Then, with the help of a text database built on DocuSky collaboration platform, a list of colour terms were generated. Further statistics and analyses focused on the frequency and categories of colour terms, revealing the conceptual colour range of Nanjing brocade. We found 73 different colour terms in the entire corpus, occurring nearly 1000 times in total. Red

and cyan are the dominant hues of Nanjing brocade, and the most commonly used colours were *dahong* (scarlet), *minghuang* (bright yellow), *chenxiang* (agilawood brown), *shiqing* (mineral blue), and so on (Fig. 2). The correlations between colour and hierarchy, gender, pattern, etc. were also explored through co-occurrence analysis. For example, the colour diversity of men's clothing of Nanjing brocade was much lower than that of women's clothing.



Fig. 2:
The distribution of Nanjing brocade colour terms: (a) Word cloud by the frequency of individual terms. (b) Hue categories proportion of all colour terms. (Colours in the chart are just indicative.)

Secondly, experiments were carried out in order to solve the problem of blurriness of dyeing knowledge. High performance liquid chromatography (HPLC) were used to detect the natural dyes in textile fragments (Pauk et al., 2014), basing on the chromatogram comparison between textile fragments and reference dyed samples (Fig. 3). Then, dyeing experiments were designed according to the information obtained from text mining, HPLC and the experience of professional dyers. Two experts participated in and conducted more than 35 experiments with 21 natural dyes. Although dealing with ancient techniques, the whole dyeing process followed the requirements of scientific experimentation. The weather, environment, materials, tools, procedures and results of the dyeing process were recorded in detail, as well as measurement data of time, amount, temperature, pH, and other dyeing conditions. Pictures and videos were also captured by digital cameras throughout the experiments. The experiment produced 434 dyed samples, all of which were measured the CIELCh value by a spectrophotometer, and each samples were drawn in the LCh colour space (Fig. 4). Comparing our natural dyeing results to synthetic dyeing swatches used nowadays, it's clear that the colour saturation of traditional Nanjing brocade should be much lower than today's products.

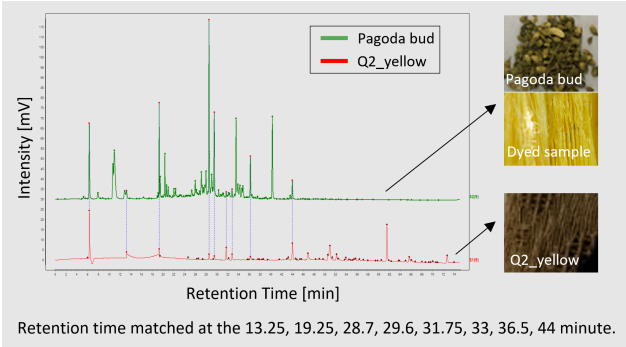


Fig. 3:
The HPLC chromatogram comparison of a yellow relic sample (red line) with reference sample dyed with pagoda bud (green line).

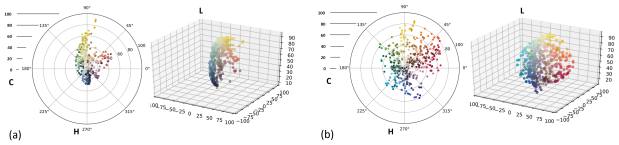


Fig. 4:
(a) The CIELCh colour space of natural dyed samples. (b) The CIELCh colour space of the Nanjing Brocade Silk Colour Swatches made with synthetic dyes.

Thirdly, the evaluation of colour perception by experts were conducted to change the subjective and blurry perception into quantitative data. Experts were asked to point out the ideal colour of each colour term in the LCh colour space. Although their perception differs a little, we could calculate the average LCh value of each colour and select a closest dyed sample.

With all the data and samples obtained, a colour specification of Nanjing brocade will be published, consisting of colour terms, colour samples (physical or digital), dyeing information and LCh values (Fig. 5). It will provide specific guidance for the heritage conservation and handicraft production of Nanjing brocade, with colour identification and communication being faster and more accurate. Besides, it is a dynamic, inclusive colour specification that will expand and update as research progresses. The methodology presented in this work, along with our literature database and natural dyes database, also provides a reference for other researchers of textile colour.

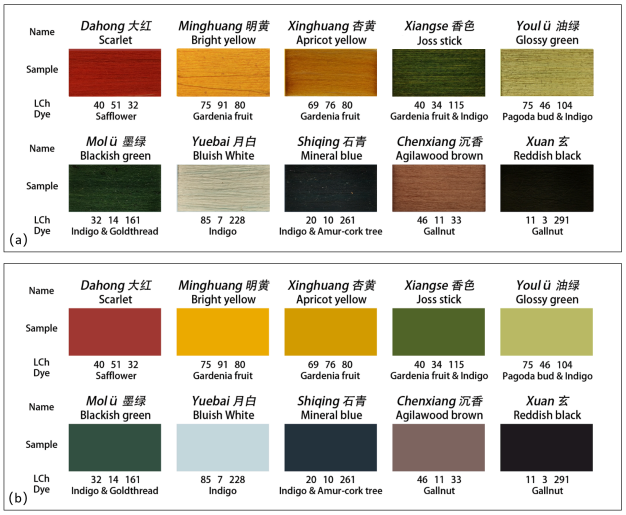


Fig. 5:
The specification of 10 typical colours of Nanjing brocade: (a) With natural dyeing silk samples. (b) With digital samples.

DocuSky Collaboration Platform. http://docusky.org.tw.

Huang, N., Wang, B., and Z. Xiao. (2003). *Nanjing Brocade of China*. Nanjing: Nanjing Press.

Kim, S. H. (2006). *The classical colors of China.* Yunnan, China: Yunnan People's Publishing House.

Li, L. (2019). New attempt on the construction of Chinese traditional colour system: Quantitative analysis and colour range exploration based on literature, relics and techniques. *The 2019 Annual Conference Proceedings on Chinese Traditional Colours.* Beijing, China: Culture and Art Publishing House.

Liu, J., Wang, Y., et al. (2020). *The Qianlong palette: The research on dyes in the 17th–19th century textiles and reconstitution of the Qing dynasty colors.* Zhejiang, China: Zhejiang University Press.

MARKUS. Text Analysis and Reading Platform. https://dh.chinese-empires.eu/beta/.

Pauk, V., Barták, P., and Lemr, K. (2014). Characterization of natural organic colourants in historical and art objects by high-performance liquid chromatography. *Journal of Separation Science,* **37**(23): 3393-3410.

Tan, A. H. (1999). Text mining: The state of the art and the challenges. *Proceedings of the pakdd 1999 workshop on knowledge discovery from advanced databases*. Vol. 8, pp. 65-70.

Digital Archives and Political Legacies: Witll the Obama Corpus Stand the Test of Time?

Losh, Elizabeth
lizlosh@wm.edu
William and Mary, United States of America

DH 2022

Digital Archives and Political Legacies: Will the Obama Corpus Stand the Test of Time?

Short Abstract:

Drawing on extensive interviews with archivists and digital strategists associated with the WhiteHouse.gov domain developed during the administration of U.S. President Barack Obama (2009-2017), as well as primary research in archival collections, this presentation explores how attempting to preserve a large and important digital archive for posterity can both solidify and undermine a political legacy.

Longer Description:

The enormous treasure trove of digital materials associated with the first Black president of the United States included text from hundreds of speeches and executive orders, video and sound files, scrapes of social media postings, and a vast repository of structured data available in various open formats. This bounty of digital information would obviously be of interest to future historians, political scientists, rhetoricians, media scholars, researchers in Black studies, and scholars in science and technology studies. Archivists at the National Archives, Internet Archive, and Library of Congress seemed to be ready to spring into action to preserve it all, beginning with harvesting materials from eyewitnesses to Obama's historic inauguration and ending with high-profile releases of tidy zipped .csv files from the POTUS and FLOTUS Twitter accounts after the transfer of power to his Republican successor had been completed. Yet behind the scenes, the story of preserving a political legacy is much more complex, particularly when issues about personal privacy, national security, the ownership of intellectual property, and platform governance come into play. This presentation draws on extensive field research done in Washington D.C. – including interviews with digital designers and strategists who managed both the front and the back end of Obama's communication infrastructure – as well as archival research in both spaces for born digital materials and traditional documentary evidence from the written record.

This presentation also discusses how political legacies of interest to digital humanists might be eroded with attention

to comparative case studies from government records of world leaders from the UK, France, and Germany. It also looks at how a digital legacy exists in a larger rhetorical and political historical context. For example, the Trump administration was quick to take down many public records made digitally available in the Obama era and even removed records about troop mobilization that had been posted during the pre-Obama Bush era of Republican political control. This context involves having digital humanists consider issues about personalization and surveillance, as the Obama administration reversed longstanding cookie policies from the Bush and Clinton White Houses. At the same time as official digital archiving practices might be noteworthy, it is also important to acknowledge the labor of DIY archivers and hacktivists, such as the maintainer of the Trump Twitter Archive. Finally, this presentation will dramatize the importance of reconstructing missing digital objects of study, such as the White House website designed for would-be president Hillary Clinton, which is not available to researchers and may become irretrievably lost without a clear mandate for preservation. This lively, provocative, and media-rich short presentation by a scholar of DH with a two-decade record of deep involvement in the field should be of interest to many attendees at DH2022.

Improving Named-Entity Recognition on Inscriptions on *ukiyo-e* prints: Towards a 'Distant Viewing' in Art History

Machotka, Ewa
ewa.machotka@su.se
Stockholm University, Sweden

Chatzipanagiotou, Marita
marita.xatzh@gmail.com
Athens University of Economics and Business, Greece.
Note: All authors have contributed equally.

Pavlopoulos, John
ioannis@dsv.su.se
Stockholm University, Sweden; Athens University of Economics and Business, Greece. Note: All authors have contributed equally.

Japanese early modern woodblock prints, so-called *ukiyo-e* or 'pictures of the floating world' produced between the seventeenth and mid-nineteenth century, are one of the most widely recognizable visual images today. Among them landscape prints remain the most popular as evidenced by the iconic "The Great Wave" designed by Katsushika Hokusai (1760-1849) and its global career (Guth 2016). However, the understanding of these images is still shaped by Western modern epistemologies that may not be well fitted for the analysis of pre-modern non-Western artefacts (Machotka 2020). The dominant Western modern concept of landscape indicates that landscape images function as representations of places (Andrews 1999) even if art is never a mirror for reality. However, this may not be the case in relation to Japanese early modern prints, which are built on poetic traditions and may have other than representational functions (Chino 2003; Machotka 2012; Shirane 2013). Therefore, to understand the relationships between images and places there is a need to look at Japanese early-modern prints afresh and artificial intelligence has a potential to aid realization of goals.

The existing discourse on Japanese landscape prints has mainly targeted case studies e.g. selected themes or artists (Clark 2001; Forrer et al. 2011; Kobayashi 2020), the approach which does not allow broad explorations of the geographical distribution of the sites depicted within the prints, their changing frequency in relation to their production context (e.g. time, location, designer) etc. Therefore, we argue that the combination of 'close reading' of the artefacts through formal and contextual analysis with so-called 'distant viewing' or macroanalysis of visual materials (Taylor and Tilton 2019) based on the idea of 'distant reading' proposed by Franco Moretti (2000) for literary studies (Gold and Klein 2016) has the potential to develop a more nuanced understanding of Japanese *ukiyo-e* prints.

Hence, with this work we propose that distant viewing can be facilitated by Natural Language Processing technologies such as Named Entity Recognition (NER). NER can be used to extract named locations from any text, including titles and other printed inscriptions on prints. Extracted locations can then allow for a digital geospatial macroanalysis of the studied prints, which is currently impossible as the artefacts form an exceptionally large and highly divergent corpus. However, although NER has the potential to improve the study of prints, the current state of the art NER tools are not successful in the identification of artwork titles (Jain and Krestel 2019). This is mainly due to the training data scarcity. Even recent cross-domain datasets only focus on domains such as politics and natural sciences (Liu et al., 2021), leaving art history aside. This problem is especially relevant for the analysis of non-Western pre-modern artefacts such as Japanese prints as inscriptions are rendered in pre-modern scripts used before the standardization of the language in the late nineteenth and twentieth centuries (Frellesvig 2012). In premodern Japanese, the Sino-Japanese characters could be used

alternately depending on their phonetic value and the same word could be written in different characters (Yada 2012). Another problem is the ambiguity inherent to the artwork inscriptions or the lack of data. Print inscriptions are not always standardized and metadata in different collections feature different information. These important issues challenge the proposed analysis.

Lee et al. (2018) were the first to show that transfer learning can lead to state-of-the-art results in NER for English patient note de-identification, by transferring learning from a large labeled dataset to a much smaller one. Following their work, we transferred a generic pre-trained Japanese Convolutional Neural Network NER model (Honnibal and Montani 2017) to the domain of art history, using a very limited training set of 100 labeled data. By using 100 (unseen) labeled data for evaluation, in a prior study (Chatzipanagiotou et al. 2021), we showed that transfer learning can assist NER in the Japanese language and in the field of art history, for the task of place name recognition in inscriptions of landscape prints. We registered an improvement of 28% in Precision, increasing it from 62% to 90%, and more than doubled F1, increasing it from 15% to 36%. We argue that the improved NER already allows distant viewing of the data and we show that there is room for further improvement. The access to data was facilitated by the database hosted at the Art Research Centre at Ritsumeikan University, Kyoto one of the leading Digital Humanities hubs in Japan and a collaborative partner of this project (http://www.arc.ritsumei.ac.jp/en/index.html).

Bibliography

Andrews, M.(1999). *Landscape and Western Art*. Oxford Univ. Press.

Chino, K.(2003). The Emergence and Development of Famous Place Painting as a Genre. *Review of Japanese Culture and Society*, 15.

Clark, T..(2001) *100 Views of Mount Fuji.* Trumbull, CT.: Weatherhill.

Forrer, M. and Suzuki, J. and Smith, H.(2011) *Hiroshige: Prints and Drawings*. Munich: Prestel.

Frellesvig, B.(2012). *A History of the Japanese Language*. Cambridge Univerity Press.

Honnibal, M. and Montani, I.(2017). SpaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. To appear.

Gold, M., and Klein, L.(2019). *Debates in the Digital Humanities*. University of Minnesota Press.

Guth, C.(2016). *Hokusai's Great Wave: Biography of a Global Icon*. Hawai'i University Press.

Jain, N. and Krestel, R.(2019). Who is mona l.? Identifying Mentions of Artworks in Historical Archives. *International Conference onTheory and Practice of Digital Libraries*. Springer, pp.115–122.

Kobayashi, F.(2020). 文政期前後の風景画入狂歌本の出版とその改題・再印:— 浮世絵風景画流行の前史として—. 浮世絵芸術 , 179: 5-19.

Lee, J. and Dernoncourt, F. and Szolovits, P.(2018). Transfer Learning for Named-Entity Recognition with Neural Networks. *The Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. pp. 4470-4473.

Liu, Z. and Yu, T. and Wenliang D. and Ji, Z. and Cahyawijaya, S. and Madotto, A. and Fung, P.(2021). *CrossNER: Evaluating Cross-Domain Named Entity Recognition." The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, pp. 13452-13460.

Machotka, E.(2020). 美術史を超えて── ヴァナキュラー・マッピングとしての日本近世風景版画. 造形のポエティカ ― 日本美術史を巡る新たな地平, ed. Hiromi N. et al. Tokyo: Seikansha.

Moretti, F.(2000). The Slaughterhouse of Literature. *Modern Language Quarterly,* 61:1.

Shirane, H.(2013) *Japan and the Culture of the four Seasons: Nature, Literature, and the Arts*. Columbia University Press.

Taylor, A. and Tilton, L.(2019). Distant Viewing: Analyzing Large Visual Corpora. *Digital Scholarship in the Humanities,* 34:1: i3–i16.

Yada, T.(2012). 矢田勉. 国語文字・表記史の研究. Tokyo: Kyūko Shoin.

Chatzipanagiotou, M. and Machotka, E. and Pavlopoulos, J.(2021). Automated Recognition of Geographical Named Entities in Titles of Ukiyo-e prints. *Digital Humanities Workshop (DHW 2021). Association for Computing Machinery*, New York, USA, 70–77.

The Expert in the Loop: Developing a Provenance Linked Open Data Management Platform

Mariani, Fabio
fabio.mariani@leuphana.de
Leuphana University Lüneburg, Germany

Provenance, as used by museums and other stakeholders in the art and cultural heritage sector, refers to the history of an artifact's ownership from its creation to its current location. Currently, provenance is stored and shared as

a text listing each ownership in chronological order, including the dates of ownership, the owners' names, the method of transfer, and the place where the objects were stored (International Foundation for Art Research, 2022) . Institutions' curators compile provenance texts following a critical analysis of historical sources, such as letters, catalogs, and inventories. Studying the ownership history of artifacts is fundamental to establishing their value and authenticity, to their attribution, and, not least, to the return of stolen objects. In addition, the information contained in these texts makes provenance particularly interesting for historical analyses of collecting, such as the study of market geography, owners' tastes, and economic and social networks. Provenance studies - the emerging discipline addressing the varied practices of provenance - operates at the intersection between digital methods and art history. Indeed, this interdisciplinary approach is effective in three areas: quantitative analysis through big data, network analysis, and spatial analysis (Jaskot, 2020).

First, however, extracting and structuring data from provenance texts is necessary to apply digital methods in provenance studies. In compliance with FAIR and Open Data principles, a Linked Open Data (LOD) publication of provenance data is highly recommended (Rother et al., forthcoming). In addition, a LOD approach allows linking provenance data between institutions, optimizing their resources, and other repositories, such as the Getty Vocabularies and the Getty Provenance Index, an index of archival inventories, sales catalogs, and dealer stock books (Davis, 2019).

Since provenance information is usually recorded in free text fields in museum collection management systems, the team of the Art Tracks project, associated with the Carnegie Museum of Art, published a software to structure provenance from scratch and generate LOD, the Elysa Tool (Newbury, 2017). Building on this state of the art project, the Provenance Lab at Leuphana University Lüneburg is developing an online data management platform. The platform aims to make the creation of provenance LOD more accessible to art historians. They can use it to generate provenance LOD either from scratch or from existing provenance texts by performing a human-in-the-loop workflow that combines artificial intelligence with the experience of the domain expert. On the one hand, it is possible to automatically extract knowledge from provenance texts through Natural Language Processing techniques. On the other hand, the intervention of domain experts is required both for data enrichment and for critical curation of the knowledge extracted by the machine within an intellectual process, defined as data literacy, in which art historians are called to participate (Klinke, 2020). This process enhances the quality of the generated data and provides feedback for machine learning. The output data's semantic structure is based on the Linked Art Data Model, an application profile of CIDOC CRM developed by the Linked Art community.

The platform front end consists of an interface where the domain expert can display, enrich, and edit the automatically extracted data and link entities to external resources suggested automatically by the machine through SPARQL queries. Each art historian's intervention is recorded to preserve the domain expert's authority, generating the data provenance of the provenance, or "The provenance of provenances," as theorized by Christian Huemer (Huemer, 2020). In addition, the platform ensures that domain experts can enrich the data at a statement level, adding references and qualifiers to handle vague, incomplete, subjective, and uncertain information (VISU data, from Latin *de visu*, "with your own eyes"). Preserving the value of VISU data is fundamental for the integrity of historical information, avoiding reductionist and objectivist bias (ter Braake et al., 2016).

Although the data management platform is currently under development, testing it has already been possible. In particular, the platform proved to be a valuable pedagogical tool during an interdisciplinary course in digital art provenance held at Leuphana University Lüneburg. Through the online interface, it is possible to actively engage students in facing the challenges arising during the digitization of historical information without requiring prior digital skills.

In presenting the data management platform, particular emphasis will be placed on the human-in-the-loop approach as a strategy to integrate the domain expert's skills within a semi-automated process. Indeed, this integration becomes a point of intersection and dialogue between Art History and Digital Humanities.

Bibliography

Braake, S. ter, Fokkens, A., Ockeloen, N. and Son, C. van (2016). Digital History: Towards New Methodologies. In Bozic, B., Mendel-Gleason, G., Debruyne, C. and O'Sullivan, D. (eds), *Computational History and Data-Driven Humanities*, vol. 482. (IFIP Advances in Information and Communication Technology). Cham: Springer International Publishing, pp. 23–32 doi: 10.1007/978-3-319-46224-0_3.

Davis, K. (2019). Old metadata in a new world: Standardizing the Getty Provenance Index for linked data. *Art Libraries Journal*, **44**(4): 162–66 doi: 10.1017/alj.2019.24.

Huemer, C. (2020). The Provenance of Provenances. In Milosch, J. and Pearce, N. (eds), *Collecting and Provenance: A Multidisciplinary Approach*. Lanham: Rowman & Littlefield Publishers, pp. 2–15.

International Foundation for Art Research (2022). Provenance Guide https://www.ifar.org/Provenance_Guide.pdf (accessed 18 April 2022).

Jaskot, P. B. (2020). Digital Methods and the Historiography of Art. In Brown, K. (ed), *The Routledge Companion to Digital Humanities and Art History*. New York: Routledge, Taylor & Francis Group, pp. 9–17.

Klinke, H. (2020). The Digital Transformation of Art History. In Brown, K. (ed), *The Routledge Companion to Digital Humanities and Art History*. (Routledge Art History and Visual Studies Companions). London: Routledge, pp. 32–42.

Newbury, D. (2017). Art Tracks: using Linked Open Data for object provenance in museums. *MW17: Museums and the Web 2017* https://mw17.mwconf.org/paper/art-tracks-using-linked-open-data-for-object-provenance-in-museums/ (accessed 18 April 2022).

Rother, L., Koss, M. and Mariani, F. (forthcoming). Taking Care of History: Toward a Politics of Provenance Linked Open Data in Museums. *Art Institute Review* (2).

Pose Clustering for Martial Arts Action Recognition: the case studies of Kata and Tai Chi

Marsocci, Valerio
valerio.marsocci@uniroma1.it
Sapienza University of Rome, Italy

Lastilla, Lorenzo
lorenzo.lastilla@uniroma1.it
Sapienza University of Rome, Italy

The history of martial arts is lost in the mists of time: from time immemorial mankind has sought methods of combat fundamental for self-defense, and not only. Just such a millenary tradition has ignited great interest in mass culture towards these techniques, often leading to false beliefs derived from myths without valid sources (Bowman, 2016).

Thus, to overcome this issue, since the end of the last century, the academy has developed an increasing interest in the martial arts, studying these disciplines from increasingly varied perspectives, such as performative arts, anthropology, history and so on. At the same time, there has been an increasing need to provide a comprehensive and consistent definition of "martial art".

According to the Japanese model (Green, 2010), "*martial arts are [...] systems that blend the physical components of combat with strategy, philosophy, tradition, or other features, thereby distinguishing them from pure physical reaction"* (Green, 2010). This definition implies that a fighting technique can be considered a martial art only when it is codified.

In this work, we try to formalize this property in mathematical terms, to make the process of studying martial arts not only qualitative, but also quantitative. In particular, we consider the enhancement of artistic image understanding through computer vision and machine learning techniques, with particular reference to the estimation and proper representation of human poses in artworks (Impett et al., 2016).

These methodologies suit well for the study of martial arts since these arts are based on encoded movements and postures of the body: in this sense, therefore, the development of tools and technologies able to facilitate the monitoring, representation and automatic analysis of human poses and movements in the execution of these arts could lead to several benefits in their practice.

Thus, we investigate the movement and the codification of the aforementioned martial arts, by performing action recognition in martial arts action sequences via human pose clustering, and following Aby Warburg's concept of *Pathosformel*, which describes the representation of emotion, movement, and passion through a repeatable visual paradigm or formula (Impett et al., 2016).

The poses are modeled as 2D skeletons and are defined as sets of 15 points connected by limbs. To perform our analysis, we make use of some of the Tai Chi and Kata sequences from the *MADS dataset*(Zhang et al., 2017).

Particularly, based on the theoretical framework established in (Marsocci et al., 2021), we carry out two series of experiments: a pose clustering to segment the whole video sequence into intervals corresponding to a given set of canonical poses; "pose spotting", based on a similarity assessment with respect to canonical poses.

Pose Clustering

We selected from the first Tai Chi progression in the MADS dataset the "Qishi" (or "Commencing Form") sequence, and we clustered the poses extracted from each frame via angular K-Medians, (with 3 clusters) based on the poses that encodes the sequence (Fig. 1). Particularly, starting from the assumption that each sequence of actions can be defined through a succession of very specific poses, we used the standard representations of these poses as predefined centroids of a partition of all the poses assumed during the sequence; as visible in Fig. 1, the obtained clusters are well ordered and easily identifiable.

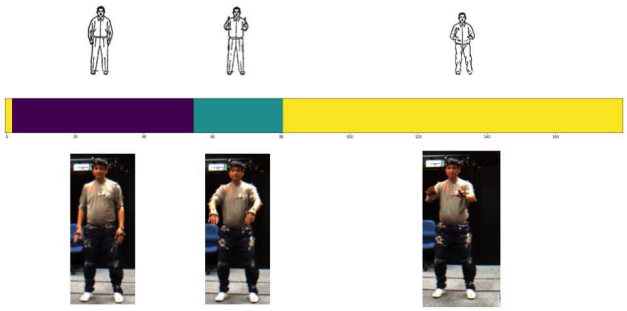


Fig. 1
Canonical poses (first line) and sequence clustered according to angular K-Medians (second line).

Similarity assessment

For the pose spotting task, we focused on the "Qishi" and "Buddha's Warrior Attendant Pounds Mortar (I)" (i.e. "Jin Gang Dao Dui (1)") sequence from the first Tai Chi progression in the MADS dataset, made of 800 frames (i.e. poses). Then, we selected two canonical poses and we computed the distance among them and the rest of the poses. As we can observe from Figs. 2 and 3, the most similar frames to the query pose can be correctly detected by setting a proper threshold on the similarity score.

Similar experiments are shown in Figs. 4 and 5 for the Kata "Fukyugata Ni" sequence (made of 600 poses). In this case, some interesting results can be outlined: the higher thresholds set for the selected poses highlight that they encode very precise movements that cannot be found in the remaining part of the sequence.



Fig. 2
Pose spotting (Tai Chi), pose 1: the frames corresponding to the query pose are correctly detected for a threshold of 0.01.

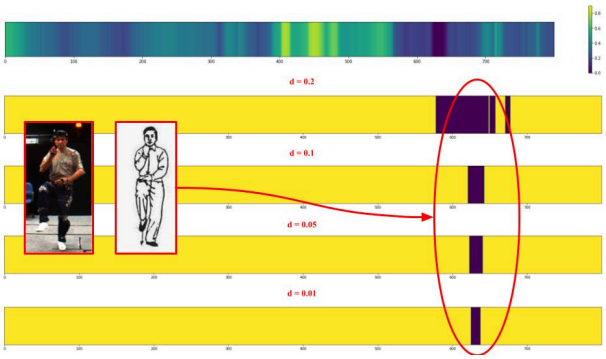


Fig. 3
Pose spotting (Tai Chi), pose 2: the frames corresponding to the query pose are correctly detected for a threshold of 0.1.

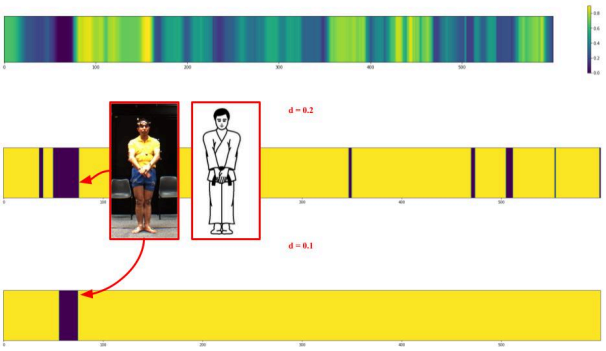


Fig. 4
Pose spotting (Kata), pose 1: the frames corresponding to the query pose are correctly detected for a threshold of 0.1.

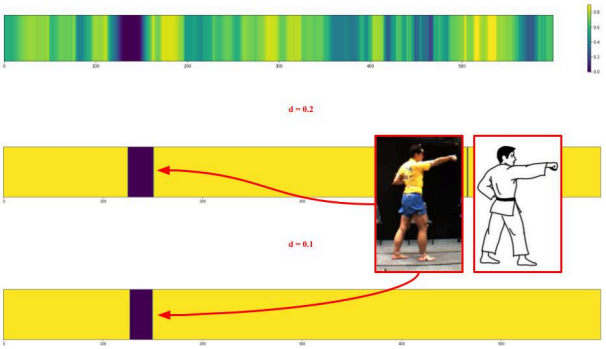


Fig. 5
Pose spotting (Kata), pose 2: the frames corresponding to the query pose are correctly detected for a threshold of 0.1.

Conclusions

Considering the aforementioned results, the downstream applications for martial arts of our framework are immediately intuitive: indeed, it could be used to compare a performed sequence to the ideal sequence of gestures, and the single poses could be compared to the coded poses too, to correct possible errors during the execution of a progression.

Bibliography

Bowman, Paul(2016). "Making martial arts history matter." *The International Journal of the History of Sport* 33.9: 915-933.

Green, Thomas A.(2010).*Martial arts of the world: an encyclopedia of history and innovation*. Vol. 2. Abc-Clio.

Impett, Leonardo, and Sabine Süsstrunk(2016). "Pose and pathosformel in Aby Warburg's bilderatlas." *ECCV*. Springer, Cham.

Marsocci, Valerio, and Lorenzo Lastilla(2021). "POSE-ID-on—A Novel Framework for Artwork Pose Clustering." *ISPRS International Journal of Geo-Information* 10.4 (2021): 257.

Zhang, Weichen, et al(2017). "Martial arts, dancing and sports dataset: A challenging stereo and multi-view dataset for 3d human pose estimation." *Image and Vision Computing* 61: 22-39.

Differences in Ideologies: Identifying Relevant Topics in Ecuadorian Presidential Speeches from 2007 – 2022

Meneses, Luis
ldmm@uvic.ca
Vancouver Island University, Canada

La República del Ecuador is a small country in South America. It is 283561 km2, making it 35 times smaller than Canada. It has a population of approximately 17 million people. The official language is Spanish, and the country uses the US Dollar as their official currency. Ecuador's economy is highly dependent on commodities, namely petroleum and agricultural products. Ecuador has a unique ecological heritage, hosting many endangered plants and animals, including those found in the Galápagos Islands.

Ecuador is currently in a state of political transition. After a series of ousted governments in the early 2000s, Rafael Correa was elected as the President of Ecuador from 15 January 2007 to 24 May 2017. Correa's political ideology was labeled as the "Socialism of the 21st Century," and it was heavily influenced by Cuba's socialism and by Fidel Castro's style of government as Prime Minister and President from 1959 to 2008. Correa's presidency was characterized by a difficult relationship with the press and was not shy of controversies. In 2020, Correa was found guilty of bribery and sentenced to serve eight years in prison. This sentence has not been executed on Correa because he left Ecuador right after his mandate ended. He is currently living in Belgium.

Correa was immediately succeeded in 2017 by Lenin Moreno, a member of his political party and one of his former Vice Presidents. As it can be expected, Moreno tried to distance himself from Correa with different policies on journalism, freedom of speech, and tackling corruption. However, the political undercurrent of Moreno's government was very similar to the one established by Correa. A political departure allegedly took place when Guillermo Lasso assumed the presidency in 2021 and distanced his government from other socialist-influenced regimes such as Cuba and Venezuela.

In this paper, I propose to continue the exploratory analysis that I have presented in previous years at the Digital Humanities Summer Institute Conference & Colloquium (Meneses, 2017) (Meneses, 2018) (Meneses, 2019) (Meneses, 2020) (Meneses, 2021). More specifically, I propose to analyze and identify the common topics in Ecuadorian Presidential speeches from fourteen years (between 2007 and 2022) by drawing inferences over those years using topic modeling. I will also contrast the Ecuadorian speeches with speeches given by Fidel Castro, whose ideology has had great influence in Ecuador and Latin America. For this purpose, I have harvested the speeches given by Correa, Moreno, Lasso, and Castro over time. These speeches were published online by the Ecuadorian Presidency (Presidencia de La República Del Ecuador) and by the Cuban Government (República de Cuba). I will address three research questions in this paper. First, as the three Ecuadorian Presidents in those years, how far apart are the political ideologies of Correa, Moreno, and Lasso? Second, can we identify any overlapping topics in their ideologies? And finally, are the ideologies of Correa, Moreno, and Lasso similar to the political ideals of Castro?

Furthermore, the Ecuadorian Presidency has systematically removed all the discourses from previous presidents, which means that most of the assets in the document corpus that I analyzed and referenced in this paper have been taken offline. Thus, this proposal contributes methods that can be used to explore sensitive situations where a document corpus is not readily available, and these situations are becoming increasingly relevant in political scenarios. The captures stored in the Internet Archive show the different sets of discourses available on the website of the Ecuadorian Presidency in 2017, 2018, and 2021 (Internet Archive, 2019).

My analysis using topic modeling has produced two core results. First, it has outlined document clusters that are clearly delimited with multiple intersecting topics; and second, it showed that increasing the topic modeling iterations increases the prominence of the overlap between the topics. Finally, the goals of this proposal are to engage in an analysis of a document corpus, identify the common trends among the four sets of discourses, and emphasize the strengths of our field, all while addressing how my research questions fit into the interdisciplinary conversation of the digital humanities.

Exploring the Correlations Between Picasso's Artworks and his Personal Contacts

Meneses, Luis
ldmm@uvic.ca
Vancouver Island University, Canada

Mallen, Enrique
edm012@shsu.edu
Sam Houston State University, United States of America

The digital humanities are at a turning point where we can digitize and analyze materials using computers with relative ease. The use of digital and computer-based methodologies can make these approaches seem innovative to researchers in the humanities, not necessarily because they are digital but because of the conclusions that can be reached. However, it is clear that we should make a departure from the traditional analyses used in other disciplines, devising our own methods to address our research questions (Orlandi, 2019).

For some time now, our research has taken us through different approaches to analyze Picasso's artistic legacy (Meneses et al., 2008a), his poetry (Meneses et al., 2008b), its semantic domains (Meneses and Mallen, 2017) and its distinctions (Meneses and Mallen, 2018). We also used computer vision to identify the graphic elements in Picasso's poetry (Meneses and Mallen, 2020). However, there are many facets of Picasso as an artist that are still unexplored.

Many authors have argued that there is a correlation between the person Picasso was with at a certain time and what he painted then, hence the so-called Olga Period, Marie-Thérèse Years, Dora Maar Years, etc. (Freeman, 1994) (Müller et al., 2002). We propose to verify if this correlation exists by examining how frequently specific individuals found in Picasso's biography are mentioned with respect to each other, and then correlate those frequencies with specific techniques, artworks, places, periods, etc. For this purpose, we will use the entries found on the extensive Online Picasso Project biographical narrative.

The Online Picasso Project (Mallen, 2018) originated from an emphasis on an interactive, digitally encoded art publication, moving away from an understanding of art criticism as predominantly stative and print-based. It consists of a complex system of interrelated databases which include both texts and images pertaining to Pablo Picasso. As a result of close collaboration between art

scholars and computer scientists, the Picasso Project has adopted an innovative architecture with three major objectives: First: to facilitate the maintenance of a large collection of artworks along with its associated critical narratives. Second, to overcome the limitations of printed art publications; and third, to provide new ways for composing, browsing, and exploring the artworks in ways not possible with printed versions. The biographical entries included in the Online Picasso Project are particularly important since they provide a historical framework that is crucial to the understanding of the artist's legacy which is tightly bound to the experiences he had in his life. Our study aims to employ this historical framework to highlight the hidden connections between Picasso's life and his artistic output.

Our methodology is based on the taxonomy-based approach that we have used to identify the semantic domains in Picasso's poetry (Meneses and Mallen, 2019), but with three differences. First, we analyze the term frequency-inverse document frequency (TF-IDF) for keywords in each biographical entry in the databases of the Online Picasso Project. TF-IDF is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. More specifically, from a list of relationships/friends/dealers, we examine who are mentioned the most and how they are related to each other. For instance, are certain dealers mentioned when a certain model is also mentioned? Second, we analyze how the frequencies in those mentions correlate with specific years/seasons/months. If so, might there be a demonstrable correlation not only with a certain model, but also with a certain dealer or collector in specific periods in his career? Finally, are certain techniques (oil, watercolor, gouache, etc) more frequent in some of those periods? If so, could they be related to the people he is in contact with? Our computational analysis aims to objectively confirm correlations that until now have been simply stipulated by art scholars.

To summarize, we propose to explore and verify the correlations between Picasso's personal contacts at certain times and his artistic output. This correlation will also provide an alternate method for exploring the existing document collection by using a curated list of keywords and metadata as an ordering criterion. In this sense, the presentation of our results draws upon our previous work that utilized interactive calendars and timelines, as it also involves utilizing appropriate user interfaces to visualize the results. In the end, the methodological contribution of our proposal is to exemplify the departure from traditional analysis while devising our own methods that are based on other disciplines (in this case statistics and computer science) to address research questions that are unique to the digital humanities.

Architectura Sinica and Collaborative Digital Database Development

Miller, Tracy
tracy.g.miller@vanderbilt.edu
Vanderbilt University, United States of America

Benda, Yuh-Fen
yuh-fen.benda@vanderbilt.edu
Vanderbilt University, United States of America

Zhuge, Jing
zhuge@seu.edu.cn
Southeast University, China

Zuo, Lala
lz2488@nyu.edu
New York University, Shanghai, China

Initially made available in 2019, Architectura Sinica (www.architecturasinica.org) is the first open-source, publicly accessible research database focusing on the timber-frame tradition of pre-modern China. Developed using the codebase of the Srophé app, and eXist DB application developed by David Michaelson (Vanderbilt University) and Dan Schwartz (Texas A&M) written in TEI, as it stands today Architectura Sinica consists of three important elements: 1. A Dynamic Site Archive of more than 140 historic (mostly religious) sites containing at least one timber-frame building thought to date from the 8 th – 13 th centuries; 2. a Chinese-English thesaurus of technical architecture terminology used by craftsmen and bureaucrats in pre-modern China; and 3. a bibliography of sources used in individual entries for historic sites, individual structures, and technical terminology. A fourth module dedicated to epigraphy preserved at these historic sites is planned for future development. Critical to this project is the display of public-domain images of historic sites and artifacts located within them. Curation of these images presents ongoing challenges as we seek to find a nimble platform for expanding our collection of digital photographs donated by researchers and students.

Over the past two years we have successfully made Architectura Sinica work as an international, collaborative, research and teaching platform. With support from the Vanderbilt University libraries, we have established a system of developing, reviewing, and publishing new data on traditional Chinese architecture in a single calendar year. This collaboration includes faculty, librarians, and

students (both undergraduate and graduate) from Vanderbilt, Southeast University 東南大學 in Nanjing, and NYU Shanghai, who work together to explore essential DH tools such as GitHub and TEI while also learning about the architectural heritage of China. We have found that this process is extremely satisfying for students and can make a seemingly esoteric topic accessible and compelling, thus sparking interest in further work in the field.

Our challenges currently revolve around the desire to keep the application open-source, while making it efficient to maintain. Securing storage for the code and data remains a problem. Our institution requires AWS hosting, and our current IT staff finds integrating our open-source XML codebase with AWS difficult. Additionally, although we currently house our high-resolution image files in FlickrPro, we have concerns about the long-term viability of using this commercial product. Yet other options, including JSTOR Forum (aka Artstor), which have a more sophisticated metadata structure, present challenges in cost and accessibility for international collaboration.

Our purpose for presenting Architectura Sinica at DH 2022 in Tokyo is two-fold. First, we would like to give a short presentation of the results of our research as an example of a collaborative DH project addressing Asian material for analysis in a global context. Second, we seek feedback on the functionality of the site. We aspire to integrate GIS data, object-based, site-specific research, and philological exploration of technical terminology into a single platform to create a useful starting point for advanced research in the history of sacred sites and the built environment in East Asia. Although our current results are satisfactory, are there ways of making the process more efficient? Are eXist DB and TEI our best tools for developing cultural heritage metadata for the future? Are there better tools available for archiving and presenting visual data in an open-source environment?

Multimodal AI support of source criticism in the humanities – work in progress

Muenster, Sander
sander.muenster@uni-jena.de
FSU Jena, Germany

Bruschke, Jonas
jonas.bruschke@uni-wuerzburg.de
JMU Wuerzburg, Germany

Hoppe, Stephan
stephan.hoppe@lmu.de
LMU Muenchen, Germany

Maiwald, Ferdinand
ferdinand.maiwald@uni-jena.de
FSU Jena, Germany

Niebling, Florian
florian.nuebling@uni-wuerzburg.de
JMU Wuerzburg, Germany

Pattee, Aaron
aaron.pattee@lmu.de
LMU Muenchen, Germany

Utescher, Ronja
ronja.utescher@uni-bielefeld.de
U. Bielefeld, Germany

Zarriess, Sina
sina.Zarriess@uni-bielefeld.de
U. Bielefeld, Germany

Introduction

The use of images, texts and objects is an essential foundation of history studies. This project funded by the *German Ministry of Education and Research* (BMBF), seeks to establish an AI-based approach towards modelling image sources and their multimodal contexts as a new technique for researchers in architectural history studies. Related questions are: How do architectural historians discover and evaluate sources? How can AI best be of service to this end?

State of the Art

The point of departure for this project is the use of sources and source criticism in history studies. This is usually led by a constructive problem-oriented approach, featuring a critical analysis of the topics and methodologies in question (Reich, 2006) and is highly experience and tacit knowledge based (Polanyi, 1966).

Language & Vision: Deep Learning (DL) prove themselves ideal for *transfer learning* at the intersection of image and language processing. For example, semantic representations such as *word* or *sentence embeddings*, which the computer learns from texts, are enriched by multimodal data such as image descriptions paired with actual visual representations (Hessel et al., 2019). However, for the extraction of multimodal information from scientific texts, it is still necessary to refine the referential connections

between text and image components (Utescher and Zarrieß, 2021).

Segmentation and Object Recognition: Photogrammetric processes deliver spatial relations between the photographs and the 3D geometries. The datasets that are developed in this method allow for the automatic segmentation (Martinovic et al., 2015, Hackel et al., 2016) of simple structures (Vosselman et al., 2004), as well as a complex objects such as buildings (Li et al., 2016, Agarwal et al., 2011).

Machine Learning (ML) is playing an ever increasing role in the segmentation of images and object recognition (Minaee et al., 2021, Jiao et al., 2019).

Research Outline

The following will provide a brief overview of the first steps in the research.

1. Identifying Research Scenarios

A series of generic scenarios were identified with the assistance of expert consultation and workshops during the preliminary investigations (Kröber, 2021, Dewitz et al., 2019), and consequently ordered by relevance and priority. Of the 20 described scenarios, the cross-media identification of object descriptions ("Which images, texts, and 3D data describe the same object?"), and the analysis of such descriptions ("How can the dating of historical image and text depictions be supported by multimodal validation using media whose dating has already been established?"), were chosen as the focal points of the research.

2. Cross-Media Classification

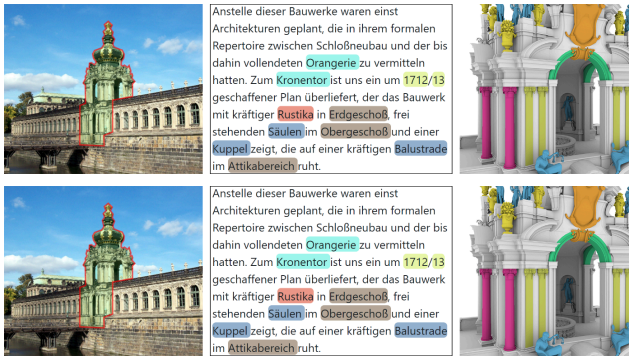


Figure 1:

Identified architectural elements using the Kronentor *of the* Zwinger *in Dresden in the photograph (left), in text (middle), and in the 3D model (right).*

A key requirement to this end, is to identify and name such cross-media elements (Fig. 1). The framework for the description of architectural elements in our project is provided by the Getty Art and Architectural Thesaurus (AAT). In our project the subgroup *architectural elements* [1] is being used. The identified elements from texts (single words or word groups), images (polygonal image details), and 3D models (individual subgroup objects) are assigned to the concept from the AAT. Different processes are necessary depending upon the source type, e.g. semantic segmentation, *Named Entity Recognition* (NER), and discourse parsing, in addition to what concerns the identification of the concepts and semantic accumulation.

3. Multimodal Data Accumulation

In a further step, various approaches are used for the accumulation and validation of multimodal data. In this way, within the 3D realm for example, 2D images can be used in relation to the 3D model in order to transfer them to the structure at hand provided by the 3D model (Niebling et al., 2018).

4. Automated classification

A current step is to investigate approaches towards automating the identification and annotation of objects. For this purpose, AI-based models will be used that are specialized on the respective modalities (3D models, images, and texts). Based on the pipeline described in (Wu et al., 2021) we currently test to enhance quality by better text identification as well as modular object retrieval for the identification of architectural structures in images (Münster et al., in print), and the transfer of this segmentation to 3D models.

Next steps

Based upon the developed demonstrator, next steps will be to cross-validate and multimodal enrich content and test those results with historians in step 1 the research scenario. It is within this area to examine the discrepancy between the requirement of large data amounts for AI models and the complexity of historical expertise can be investigated and evaluate, how existing AI models can be employed within the field of architectural history research and criticism.

Bibliography

AGARWAL, S., FURUKAWA, Y., SNAVELY, N., SIMON, I., CURLESS, B., SEITZ, S. M. & SZELISKI, R. 2011. Building rome in a day. *Communications of the ACM,* 54 **,** 105.

DEWITZ, L., KRÖBER, C., MESSEMER, H., MAIWALD, F., MÜNSTER, S., BRUSCHKE, J. & NIEBLING, F. 2019. Historical Photos and Visualizations: Potential for Research. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* XLII-2/W15 **,** 405–412.

HACKEL, T., WEGNER, J. D. & SCHINDLER, K. 2016. Fast semantic segmentation of 3D point clouds with strongly varying density. *ISPRS Annals,* 3 **,** 177–184.

HESSEL, J., LEE, L. & MIMNO, D. Unsupervised Discovery of Multimodal Links in Multi-image, Multi-sentence Documents. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019.

JIAO, L., ZHANG, F., LIU, F., YANG, S., LI, L., FENG, Z. & QU, R. 2019. A Survey of Deep Learning-Based Object Detection. *IEEE Access,* 7 **,** 128837-128868.

KRÖBER, C. 2021. German Art History Students' use of Digital Repositories: an Insight *Papers Proceedings, Diversity, Divergence, Dialogue.* Cham: Springer LNCS.

LI, M., NAN, L., SMITH, N. & WONKA, P. 2016. Reconstructing building mass models from UAV images. *Computers & Graphics,* 54 **,** 84-93.

MARTINOVIC, A., KNOPP, J., RIEMENSCHNEIDER, H. & VAN GOOL, L. 3d all the way: Semantic segmentation of urban scenes from start to end in 3d. IEEE Computer Vision & Pattern Recognition, 2015. 4456–4465.

MINAEE, S., BOYKOV, Y. Y., PORIKLI, F., PLAZA, A. J., KEHTARNAVAZ, N. & TERZOPOULOS, D. 2021. Image Segmentation Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **,** 1-1.

MÜNSTER, S., LEHMANN, C., LAZARIV, T., MAIWALD, F. & KARSTEN, S. in print. Toward an Automated Processing Pipeline for a Browser-based, City-scale Mobile 4D VR Application Based on Historical Images. *In:* NIEBLING, F. & MÜNSTER, S. (eds.) *Proceedings of the 2nd UHDL Workshop.* Cham: Springer CCIS.

NIEBLING, F., MAIWALD, F., MÜNSTER, S., BRUSCHKE, J. & HENZE, F. Accessing Urban History by Historical Photographs. 2018 3rd Digital Heritage International Congress (DigitalHERITAGE) held jointly with 2018 24th International Conference on Virtual Systems & Multimedia (VSMM 2018), 2018 San Francisco. 1-8.

POLANYI, M. 1966. *The tacit dimension,* Chicago, University of Chicago Press.

REICH, K. 2006. Konstruktivistische Ansätze in den Sozial- und Kulturwissenschaften. *Konstruktivistische Didaktik: Lehr-und Studienbuch mit Methodenpool.* Beltz.

UTESCHER, R. & ZARRIEß, S. What Did This Castle Look like before? Exploring Referential Relations in Naturally Occurring Multimodal Texts

Proceedings of the Third Workshop on Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN), 2021.

VOSSELMAN, G., GORTE, B. G., SITHOLE, G. & RABBANI, T. 2004. Recognising structure in laser scanner point clouds. *ISPRS Archives,* 46 **,** 33-38.

WU, X., AVERBUCH-ELOR, H., SUN, J. & SNAVELY, N. 2021. *Towers of Babel: Combining Images, Language, and 3D Geometry for Learning Multimodal Vision*.

Notes

1. http://vocab.getty.edu/aat/300000885, 15.07.2021.

Semiotically Unbound: People, Pedagogies and Dialogues - Beginning to do Digital Humanities at Shanghai University.

Murphy, Orla
o.murphy@ucc.ie
University College Cork, Ireland

Xiao, Shuang
shuang.xiao@umail.ucc.ie
Shanghai University, China

This paper presents a bottom-up, person-centred paradigm for building and integrating Digital Humanities (DH) with Shanghai University (SHU). Based on five years of Digital Humanities exchange between SHU and University College Cork (UCC) it considers the nuanced negotiations it takes in practice to build relationships and create mutual trust. It is a reflection on a serendipitous encounter that developed into a strategic approach to research dialogue, with inclusive, responsive co-design centred on the SHU Summer School. It has changed since its first iteration in 2013, from 3-5 people initially enrolled to 3,000 people watching live online via BiliBili, live broadcast has amplified the impact of the 'Introduction to Digital Humanities'.

The ten-lesson course emerged as one offering in an international summer course program, delivered over one

week. An introduction to the emergence of DH, some methods, and tools, and practicing HTML/XML for BA students. The content was accessible and encouraging for students - empowering them to code, to create, and present work in an engaging approach. In the face-to-face course the professor directly engages with and gauges students' progress caring about students' learning, live in class. The online course has more audience, recordings are stored and shared widely. Student feedback is positive, it is the beginning of their DH journeys.

Digital Humanities is happening on the Chinese mainland. Scholars are engaging with Digital Humanities at all levels, encouraging international staff and student exchange, and promoting Digital Humanities as a new way of engaging with arts and humanities contexts through summer schools, training courses and online pedagogies. There are structural signs within the Chinese Academy that DH is gaining traction. This paper is a lens through which to view a successful local integration that is surviving across continents through the pandemic.

Chinese debates in the Digital Humanities echo the concerns of those of us already doing DH for decades. What does this tell us in terms of how we are growing as a field of scholarly endeavour? They are asking:

What is the relationship between public knowledge service institutions or GLAM sector, and academic research?

Is a system for evaluating Digital Humanities research and promotion possible?

What does Digital Humanities education look like, how do we start?

What is Digital Humanities pedagogy?

How does China grapple with a predominantly anglophone mainstream Digital Humanities?

These practical concerns are mirrored in philosophical questions of the nature of digital art and culture, and critical theory. Chinese Digital Humanities holds a mirror to Digital Humanities more broadly, have we answered these questions where Digital Humanities has been established for decades? Do we have answers to these questions that will encourage early career researchers in China to undertake Digital Humanities work? Can we provide them with a favourable development environment and a feasible development path? These questions are faced at SHU and UCC together.

We are all concerned with evaluation, with sustainability and with impact – what might Digital Humanities do for humanities researchers in China? What does DH bring to the students? Culturally distinct in terms of pedagogy and the practice of research, this paper marks a way point for this encounter – it shares what we have learned from each other and valorises shared humanistic concerns and values. It is not a digital method or tool. It is doing Digital Humanities whilst semiotically unbound, responding to Asian diversity and a sharply focused exploration of how DH is developing in one major university SHU in conversation with UCC.

It differs from top-down approaches imposed by international deans and offers another perspective for growth, mutuality, and directions for Digital Humanities outside the anglophone world. It is a work in progress. It initiates and invites further conversations about the dynamics of exchange.

Bibliography

Gao, Jin et al. (2018) "What do we write about in the Digital Humanities? A Comparative Study of Chinese and English Publications." Discovery Open Access Repository, UCL.

(accessed 8 December 2021).

Wang, Chen et al. (2021) "The Compilation of Speeches on 'Evaluating of Digital Humanities Scholarship: Definitional and Standards' Conference", *Journal of Digital Humanities,* （《数字人文》

1: 1-57.

Xie, Liping. (2020) "Digital Humanities Dialogue and Controversy Through the Comparative Insight: A Seminar Review of 'Rethinking Digital Humanities Through the Comparative Insight' in Nanjing University", *Journal of Digital Humanities*, **3**:151-168.

Xiao, Shuang. (2020) "Getting Started in Digital Humanities: A Summary of 'Introduction to Digital Humanities' Course from University College Cork," *Journal of Digital Humanities,* **3**:175-179.

VedaWeb 2.0: Towards a Collaborative Workspace for Indo-Aryan Texts

Neuefeind, Claes
c.neuefeind@uni-koeln.de
University of Cologne, Germany

Kölligan, Daniel
daniel.koelligan@uni-wuerzburg.de
University of Würzburg, Germany

Reinöhl, Uta
uta.reinoehl@linguistik.uni-freiburg.de
University of Freiburg, Germany

Sahle, Patrick
sahle@uni-wuppertal.de

University of Wuppertal, Germany

Introduction

VedaWeb, developed in the course of a project funded by the German Research Foundation (DFG) from 2017-2021, is a web-based platform for linguistic and philological work with Old Indo-Aryan texts (https://vedaweb.uni-koeln.de). In this contribution, we report on the key features of the VedaWeb platform and give a prospect on the recently started follow-up project VedaWeb 2.0 (2022-2025), which aims at transforming VedaWeb from a locally developed and curated platform into a collaborative workspace which will be shaped by the needs of the community of Indo-Aryanist researchers as a whole.

The VedaWeb project

The VedaWeb platform was developed on the basis of the Rigveda (about 160.000 words) as the pilot text, supplemented by multiple translations and linguistic annotations, lemma- as well as word form-based links to dictionary entries, alongside a tailored search engine based on elasticsearch. [1] The Rigveda was annotated by researchers of the University of Zurich with rich, albeit partially incomplete morphological glosses. Gaps in the annotations were filled in and existing ones stream-lined by the VedaWeb project team. Moreover, novel information was added (e.g. morphological annotations of verb classes, verbal stem formations such as desideratives, nominal forms such as comparatives and superlatives, word classes such as local particles, etc). Additional translations, metrical information, and access to APIs for Sanskrit dictionaries (Mondaca et al. 2019a/2019b) [2] were integrated. All data layers have been modeled and encoded in TEI-P5 [3] and a state-of-the-art web application [4] has been developed for accessing data offered via APIs. [5]

There are several web portals making ancient Indo-Aryan texts available such as TITUS (Thesaurus Indogermanischer Text- und Sprachmaterialien, University of Frankfurt [6]), GRETIL (Göttingen Register of Electronic Texts in Indian Languages, University of Göttingen [7]), DCS (Digital Corpus of Sanskrit, University of Heidelberg [8]) and SARIT (Search and Retrieval of Indic Texts [9]). By providing a combination of several text versions, translations, linguistic analyses and combined searchability across morphological, syntactic, lexical and metrical features, VedaWeb by far exceeds the scope of existing platforms and has already significantly improved the digital presence of Old Indo-Aryan texts.

VedaWeb 2.0

Primary goal of the follow-up project is to develop the VedaWeb platform into a collaborative workspace to open it up to more texts and to further enrichment through annotations and analyses. Particularly in view of the widespread traditional practices of work in this field that hinder local and international collaboration (e.g., excel files, paper, or similar approaches), the further development of VedaWeb promises substantial advances for anyone working with Indo-Aryan texts, including linguists, philologists, cultural anthropologists, and researchers in philosophical and religious studies.

Furthermore, we will add several new textual and lexical resources. This includes the Śaunaka version of the Atharvaveda, the second oldest Vedic text, and the early Vedic prose text Aitareya Brahmana. These texts and their nearly complete morphological annotation will be provided by our project partners at the University of Zurich (Paul Widmer and Oliver Hellwig). Our cooperation partner J.S. Kim (Würzburg) will provide an *index verborum* to the Atharvaveda. [10] In addition, parts of the Maitrayani Samhita, Jaiminiya Brahmana, and the Śatapatha Brahmana will be added. As for lexical resources, we will integrate several additional dictionaries that are part of the Cologne Digital Sanskrit Dictionaries. [11] These have been modeled in TEI-P5 and OntoLex-Lemon (Mondaca/Rau 2020), and are available via the C-SALT APIs for Sanskrit Dictionaries. [12]

Additionally, we will include audio recordings of recitations of the Rigveda and Atharvaveda, made available through the National Library of Denmark. [13] These recitations stand in the tradition of the uninterrupted oral transmission of the early sacred texts of Hinduism since their conception roughly three millennia ago. Supported by the Language Archive Cologne [14], the recordings will be segmented and made available for online listening, aligned with the corresponding hymns, stanzas and verses.

Summary/Outlook

VedaWeb is already a well-established platform for linguistic and philological work with Old Indo-Aryan texts. VedaWeb 2.0 will further boost research on Old Indo-Aryan languages in linguistics, philology, and beyond. As a collaborative research platform, VedaWeb 2.0 will integrate several additional Indo-Aryan texts which can be enriched by users with diverse types of annotations, multiple translations, and links to dictionary entries. Data layers in all possible combinations and in a variety of formats will be exportable. In addition, a powerful search engine will allow combined queries across all data layers, including similarity

searches. Above and beyond the already sizable number of users of VedaWeb 1.0, we expect a further increase in participants working and networking on the platform.

Bibliography

Kiss, B., Kölligan, D., Mondaca, F., Neuefeind, C., Reinöhl, U., Sahle, P. (2019): *It Takes a Village: Co-developing VedaWeb, a Digital Research Platform for Old Indo-Aryan Texts.* In: Steven Krauwer und Darja Fišer (eds), *TwinTalks at DHN 2019 – Understanding Collaboration in Digital Humanities.* Kopenhagen, 2019.

Mondaca, F., Rau, F., Neuefeind, C., Kiss, B., Kölligan, D., Reinöhl, U., Sahle, P. (2019a): *C-SALT APIs - Connecting and Exposing Heterogeneous Language Resources.* In: Book of Abstracts of the Digital Humanities Conference 2019 (DH2019) 09.07-12.07.2019. Utrecht, Netherlands.

Mondaca, F., Schildkamp, P., Rau, F. (2019b): *Introducing Kosh, a Framework for Creating and Maintaining APIs for Lexical Data.* In: Electronic Lexicography in the 21st Century. Proceedings of the eLex 2019 Conference, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., pp. 907–921.

Mondaca, F. and Rau, F. (2020): *Transforming the Cologne Digital Sanskrit Dictionaries into Ontolex-Lemon. Proceedings of the 7th Workshop on Linked Data in Linguistics: Building Tools and Infrastructure at LREC 2020.* 11–16 May 2020. Marseille, France, pp. 11-14.

Reinöhl, U., Kölligan, D., Kiss, B., Mondaca, F., Neuefeind, C., Sahle, P. (2018): *VedaWeb – eine webbasierte Plattform für die Erforschung altindischer Texte.* In: Book of Abstracts der 5. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd 2018), Köln 26.2.–2.3.2018, pp. 485–486.

Notes

1. https://vedaweb.uni-koeln.de/rigveda/
2. https://cceh.github.io/c-salt_sanskrit_data
3. https://github.com/cceh/c-salt_vedaweb_tei
4. https://github.com/vedawebplatform/vedaweb
5. See also Kiss et al. (2019) for reflections on collaboration within and beyond the project team.
6. http://titus.uni-frankfurt.de/indexe.htm
7. http://gretil.sub.uni-goettingen.de/gretil.html
8. http://www.sanskrit-linguistics.org/dcs/
9. http://sarit.indology.info/
10. https://www.phil.uni-wuerzburg.de/fileadmin/04080400/2019/Atharvaveda_Index_verborum_10.2019.pdf
11. https://www.sanskrit-lexicon.uni-koeln.de
12. https://cceh.github.io/c-salt_sanskrit_data
13. http://www.kb.dk/en
14. https://lac.uni-koeln.de/

Quantifying Representations of Asian Identity in 21st-century Anglophone Fiction for Young Readers

Nomura, Nichole Misako
nnomura@stanford.edu
Stanford University

Dombrowski, Quinn
qad@stanford.edu
Stanford University

Since 2019's "Own Voices" campaign, publishers of youth literature in the United States have been actively seeking out authors who embody "diversity" along the axes of race, ethnicity, and sexual orientation, and who write characters that reflect some of those experiences. While there has been some backlash to the tendency to reduce complex, and sometimes private, author identities to a set of public marketing labels, new fiction released for middle-school and YA readers in 2021 is rhetorically positioned against an implicit monolith of straight, white, socioeconomically comfortable narratives thought to dominate youth literature throughout the 90s and 2000s.

This short paper uses a sample from the Young Readers Database of Literature (YRDL) to examine the portrayal of Asian diversity in American middle-school and YA novels written in English between 2000 and 2020. This new corpus contains over 22,000 Anglophone fiction novels for young readers published in the 20th and 21st century. This is substantially more comprehensive than the corpora underpinning earlier computational work on Anglophone YA (e.g. 200 works make up the YA corpus referenced in Piper 2018) and opens up new possibilities for quantitative research on young readers' fiction that reaches beyond award-winners and immediately-recognizable titles. For this study, we sampled 250 texts from each year of the 20-year period, for a total of 5,000.

In this study, we track the frequency of both proper nouns and adjectives referring to Asia and individual Asian countries (e.g. "Japan" and "Japanese"). We also take a particularly close look at the compound adjective [X]-American, a common means of marking an "otherness" from the implied white, Western European default that does

not rise to the level of "foreignness." We seek to quantify and map explicit mentions of national identity and nations to help sustain and continue the ongoing conversation about who is represented, how, in literature for young readers.

We find that approximately 60% of the books in this sample include at least one Asian country name or adjective. In the majority of books, there are a small number of passing references; the median number of references is 4. Places and adjectives followed the same trends; Chinese, Indian, Russian, and Japanese were most common; "Asian" was used more frequently than the remaining 34 adjectives. Comparing adjective use over the 20-year time period, the only adjective that showed a significant pattern of change was "Korean," which has increased between the mid-2010s and 2020, corresponding to the broader rise in the popularity of Korean culture worldwide.

Given the high frequency of references to "Chinese" and "Indian," two of the most common origins for Asian-Americans, we explored the correspondence between demographic representation and references in these texts. We had particularly noteworthy findings for "Japanese," which occurred with 2.7x the frequency of what we would predict based on the number of people who identify as Japanese-American. "Vietnamese" and "Filipino" were dramatically underrepresented, at 16% and 2% of the predicted value based on demographics.

Using spaCy NLP dependency parsing, we examined the nouns that the adjectives were modifying in these texts. The most common noun was "food," which half the time was modified by "Chinese," but a wide variety of adjectives appeared in this culinary context, including Thai, Korean, Asian, Cambodian, Egyptian, Turkish, and Malaysian. "Restaurant," the fourth most-common noun, showed similar trends. "Asian" usually modified generic words for people (e.g. girl, man, boy, guy). Some groups of nouns appeared predominantly or exclusively with a small number of adjectives: military nouns like "soldiers" and "army" were primarily associated with "Japanese," "Russian," and "Israeli." "Government" was most often modified by "Chinese" and "Russian." A few adjectives had unique groups of associated nouns: for "Egyptian," most of the references are historical and specific: god, hieroglyphics, mummy, exhibit, tomb. "Armenian" was most often connected to "war" and "genocide."

Parsing and counting these references allows us to begin to construct a connotative and denotative map of explicit Asian identity in Anglophone literature for young readers. Useful as numbers, these references also lay the groundwork for future qualitative analysis and quantitative clustering, as we seek to identify texts that respond to or challenge existing discourse conventions.

Bibliography

Piper, Andrew. (2018). *Enumerations.* Chicago: University of Chicago Press.

An International Perspective on Creating an Army of Hacker-Scholars

Öhman, Emily Sofi
ohman@waseda.jp
Waseda University, Japan

With the increased demand in industry for people with programming skills and the visibility of Digital Humanities in academia, there has been a rise in humanities and other non-computer science students looking to learn programming (Hughes, 2020). However, programming course availability and quality at humanities departments vary greatly (see e.g Abrahams, 2010; Lyon & Magana, 2020) and students risk losing motivation if they cannot see the connection to their own studies with numerical rather than textual content and practical applications (Forte & Guzdial, 2005; Ramsay, 2012; Kokensbarger et al., 2018; Öhman, 2019).

Some research already exists on what makes a programming course successful in terms of high pass rates (see e.g. Nikula et al., 2011), however, in this paper we focus on examining what practices convince students themselves that a course was useful as a measure of success. In order to examine this aspect and to try to create a few guidelines for successful Python courses specifically for students of humanities-related subjects, we examined student course feedback from three courses: an undergraduate course *Python Programming for Digital Humanities*(~150 students, early 20s, online) at Waseda University, Japan and the *NLP for Linguists* and *Working with Text* courses, partially based on the *Applied Language Technology* online tutorials (Hiippala, 2021), at the University of Helsinki, Finland (~30 students each, early to late 20s, hybrid style) that are open both for undergraduate and graduate students. The main contents of these courses are fairly similar and so are both the complaints and praise. We compare difficulties students face, explore what keeps them motivated, and analyze their feedback holistically while looking for common denominators of what works:

"These courses have been the high point of the spring semester, and all of the exercises have been motivating and satisfying to complete." (Student. Finland)

As is common with humanities-focused programming courses there is a significant skill gap between students (see

e.g. Öhman, 2019) and this too shows in the evaluations. In 2022 the question "What was the level of this week's content?" was asked of students after their first week of the "Python Programming for Digital Humanities" course. With 107 respondents, the distribution of responses shows normal distribution with most students (69%) feeling that the level was just right, and the rest almost evenly split between "I struggled a bit" and "Too easy" (Table 1).

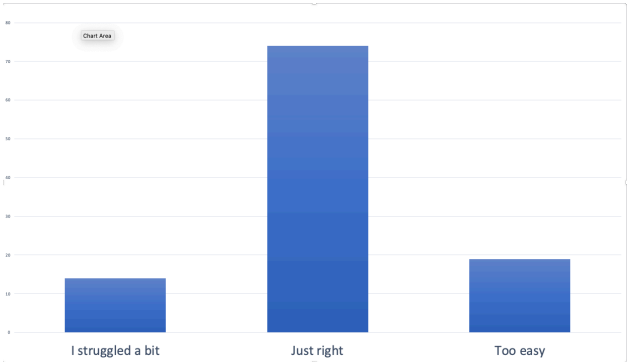


Table 1.
Results of poll regarding content difficulty

In the Finnish data, the question "Was the level of the course appropriate?" the answers have the greatest spread as well, suggesting that ideally perhaps students should either be placed in different groups based on initial skill level or that some students could really benefit from a pre-introduction course or extra tutoring where they could gain confidence in the most basic of programming concepts. This is especially true in the light that there does not seem to be a correlation between struggling students and below average final grades as long as the students do not drop out. A correlation matrix of evaluation questions and results (Figure 1.) show that there are high correlations between how students experience the speed of the course and the workload and the level of the course. If one of these parameters are adjusted it will likely affect the others, e.g., if the speed that new topics are introduced is slowed down, the workload will feel more manageable and the contents easier to digest.

We also recommend that programming courses such as these are best balanced by using scaffolding methods to keep students in the zone of proximal development (Chaiklin et al, 2003; Vygotsky, 1987). In practice this means telling the students the outline of what they are going to learn first to enable independent learning at the students' own pace later and not making the contents too easy, but also providing students with the tools to do their own trouble-shooting and advanced learning by introducing StackOverflow and other such tools early on. Additionally, the content should be humanities focused and practical. This could mean introducing the NLTK library (Bird & Loper, 2004) right after teaching the basics (data types & loops) as a way to demonstrate usefulness.

Overall, student satisfaction ratings are very high for both courses as are enrollment numbers. Ramsay (2012) referred to teaching humanities students programming as raising an army of hacker-scholars, and it certainly seems that the interest is there from the student side once they get past the initial hurdle of enrolling in the course.

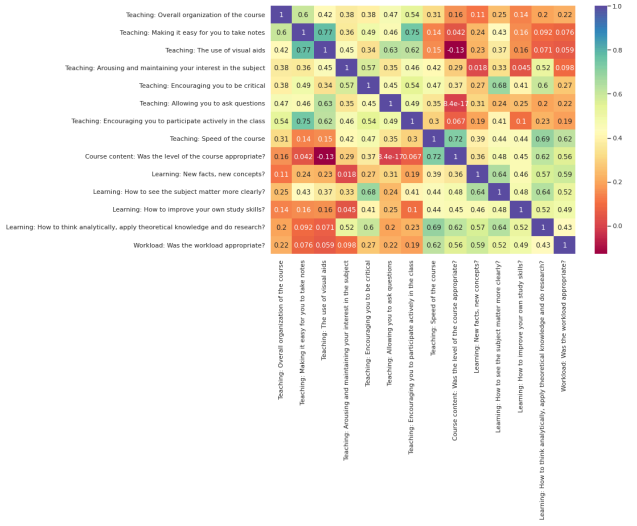


Figure 1.
Correlation matrix of feedback responses

Bibliography

Abrahams, D. A. (2010). Technology adoption in higher education: A framework for identifying and prioritizing issues and barriers to the adoption of instructional technology. *Journal of Applied Research in Higher Education* 2, 2, 34–49.

Bird, S. G., & Loper, E. (2004). *NLTK: the natural language toolkit*. Association for Computational Linguistics.

Chaiklin, S. (2003). The zone of proximal development in Vygotsky's analysis of learning and instruction. Vygotsky's educational theory in cultural context 1, 39–64.

Forte, A., and Guzdial, M. (2005). Motivation and nonmajors in computer science: Identifying discrete audiences for introductory courses. *IEEE Transactions on Education* 48:2, 248–253.

Hiippala, T. (2021). Applied Language Technology: NLP for the Humanities. *In Proceedings of the Fifth Workshop on Teaching NLP* (pp. 46-48).

Hughes, O. (2020). *Developer training sees spike in demand as more people learn to cod* e. TechRepublic. Retrieved November 29, 2021, from https://

www.techrepublic.com/article/the-economic-outlook-is-uncertain-so-more-people-want-to-become-developers/

Kokensparger, B., and Peyou, W. (2018). Programming for the humanities: A whirlwind tour of assignments. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education,* SIGCSE'18, ACM, pp. 1050–1050.

Lyon, J.A., and J. Magana, A. (2020). *Computational thinking in higher education: A review of the literature*. Computer Applications in Engineering Education.

Öhman, E.S. (2019). Teaching Computational Methods to Humanities Students. In *Digital Humanities in the Nordic Countries Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*. CEUR-WS.org.

Nikula, U., Gotel, O. and Kasurinen, J. (2011). A motivation guided holistic rehabilitation of the first programming course. *ACM Transactions on Computing Education* (TOCE), 11(4), pp.1-38.

Ramsay, S. (2012). Programming with humanists: Reflections on raising an army of hacker-scholars in the digital humanities. *Digital Humanities Pedagogy: Practices, Principles, and Politics*, 217–41.

Vygotsky, L. (1987). Zone of proximal development. *Mind in society: The development of higher psychological processes* 5291, 157.

Perspectives on the Future of Digital Editions & Publishing

O'Sullivan, James
james.osullivan@ucc.ie
University College Cork

Pidd, Michael
m.pidd@sheffield.ac.uk
Digital Humanities Institute (DHI), University of Sheffield

Murphy, Órla
o.murphy@ucc.ie
University College Cork

Wessels, Bridgette
bridgette.wessels@glasgow.ac.uk
University of Glasgow

Acknowledgements:
Postdoctoral Research Fellow
Michael Kurzmeier
mkurzmeier@ucc.ie
University College Cork

The digital scholarly edition remains central to the intellectual practices of the arts and humanities, and yet, the fundamentals of their form and structure remain unchanged by the affordances of computers. The edition is often the version of the primary source that is most immediate, accessible, and informative to scholars and students alike, and so it is vital that we invest in further enhancing that dialogue and enable researchers to establish the methods and principles for developing the scholarly digital editions of the future.

C21 Editions [1] is a three-year international collaboration jointly funded by the Arts & Humanities Research Council (AH/W001489/1) and Irish Research Council (IRC/W001489/1). The aim of the project is to investigate and advance the practices of digital scholarly editing and publishing by researching and prototyping data standards that accommodate born-digital content such as social media, while also further integrating the "curatorial" and "statistical" aspects of DH [2] by examining how computer-assisted analytical methods can be embedded into edition making and publishing.

The first phase of *C21 Editions* project engaged in semi-structured interviews with an extensive group of 50 experts and stakeholders from a range of relevant disciplines, including digital scholarly editing, digital publishing, archiving and preservation, interface design, and creative practice. This paper present the results of a thematic analysis of those interviews, providing a comprehensive overview of how many of the field's most prominent theorists and practitioners view the present state of digital scholarly editing and publishing, and how the technical systems and models which facilitate the making of digital editions and public resources might and should develop into the future. These findings will further serve as a vital compendium and roadmap for the future of digital scholarly editing, comprising perspectives by those positioned to realise any such future.

Overall, *C21 Editions* is intended to operate as a response to Joris van Zundert, who calls on theorists and practitioners to "intensify the methodological discourse" necessary to "implement a form of hypertext that truly represents textual fluidity and text relations in a scholarly viable and computational tractable manner" (2016, 106). "Without that dialogue," he warns, "we relegate the raison d'être for the digital scholarly edition to that of a mere medium shift, we limit its expressiveness to that of print text, and we fail to explore the computational potential for digital text representation, analysis and interaction." This dialogue has begun in earnest (Driscoll and Pierazzo 2016; Boot et al. 2017), but a previous survey on the expectations and use of digital editions found that user needs are seldom

satisfied by such resources (Franzini, Terras, and Mahony 2019). Initial findings from the extensive qualitative data being analysed as part of *C21 Editions* suggests that this dissatisfaction persists, and despite the considerable amount of effort going into the development of digital editions, [3] there remains a disconnect between digital cultural resources and the needs for their users. The findings of this thematic analysis builds on existing research through providing extensive insight into how it is that those tools and methods that dominate digital scholarly editing and publishing have not advanced considerably since van Zundert's statements in 2016.

Through the many expert perspectives that *C21 Editions* has gathered and analysed, this paper shows how key stakeholders believe digital editing and publishing have advanced pre-digital practices, where the digital has failed to realise its potential, and how we might envision future conditions.

[1] See c21editions.org.

[2] See Bode (2019).

[3] As evidenced by "A Catalogue of Digital Editions" (Franzini, Terras, and Mahony 2016) or dig-ed-cat.acdh.oeaw.ac.at/.

Bibliography

Bode, K. (2019). Computational Literary Studies: Participant Forum Responses, Day 2, *In the Moment* https://critinq.wordpress.com/2019/04/02/computational-literary-studies-participant-forum-responses-day-2-3/ (accessed 27 July 2020).

Boot, P., Cappellotto, A., Dillen, W., Fischer, F., Kelly, A., Mertgens, A., Sichani, A.-M., Spadini, E. and Hulle, D. van (eds). (2017). *Advances in Digital Scholarly Editing*. Sidestone Press.

Driscoll, M. J. and Pierazzo, E. (eds). (2016). *Digital Scholarly Editing: Theories and Practices*. Open Book Publishers, doi:10.11647/OBP.0095. https://www.openbookpublishers.com/product/483 (accessed 8 December 2021).

Franzini, G., Terras, M. and Mahony, S. (2016). A Catalogue of Digital Editions. In Driscoll, M. J. and Pierazzo, E. (eds), *Theories and Practices: Digital Scholarly Editing.* Cambridge: Open Book Publishers, pp. 161–82.

Franzini, G., Terras, M. and Mahony, S. (2019). Digital Editions of Text: Surveying User Requirements in the Digital Humanities. *Journal on Computing and Cultural Heritage*, 12(1): 1:1-1:23, doi:10.1145/3230671.

van Zundert, J. (2016). Barely Beyond the Book?. In Driscoll, M. J. and Pierazzo, E. (eds), *Theories and Practices: Digital Scholarly Editing*. Cambridge: Open Book Publishers, pp. 83–106.

Towards a Global Analysis of Changes in Shape over Time based on Digitised Artefacts: The East-Asian Perspective

Pala, Giovanni
giovanni.pala@magd.ox.ac.uk
University of Oxford, United Kingdom

Costiner, Lisandra
lia.costiner@history.ox.ac.uk
University of Oxford, United Kingdom

Liu, Yidan
Y.Liu-237@sms.ed.ac.uk
University of Edinburgh, United Kingdom

Wang, Shuofei
Kwwn34@durham.ac.uk
University of Durham, United Kingdom

This study tests the use of a novel computational approach, one that analyses changes in shape of historical artefacts across time, in a new context. Previously developed by the authors and tested upon Western art, in particular ancient Greek pottery, this methodology is here applied for the first time to East-Asian art, in particular Chinese vases [1]. The East-Asian perspective is crucial in understanding the adaptability of the approach to different geographical regions and time periods, contributing to the construction of a global history of shape evolution and design progression over time.

The study of shapes and styles as embodying the cultural concerns of a particular historical moment has been at the center of several disciplines including art history and archaeology. It has captured the interest of scholars since the eighteenth century when Johann Joachim Winckelmann devised his categorisation of style, focusing particularly on Greek and Roman art [2]. In more recent times, George Kubler proposed new ways of historical sequencing of form based on continuous change across time [3]. In Chinese art, surveys of the development of pottery over time have also been conducted, most recently by Ye Zhemin叶喆民 [4].

The current research inscribes itself within this intellectual tradition yet propose a new way of quantifying changes in shape and exploring connections between

objects: a computational technique. A few studies have attempted to move in this direction although they have been restricted by the technology, and materials, namely photographs [5, 6]. This paper employs a new methodology and material, 3D scans of historical artefacts, therefore providing one of the first case studies of corpus research on 3D digitised objects.

The approach has been tested on a case study of four Chinese vases of the Beaker type, deriving from late Ming and early Qing Dynasties (1620-1683). These objects are held in the Ashmolean Museum in Oxford, U.K., under ascension numbers EA1978.799, EA1978.798, EA1971.22, and EA1978.1903. These were chosen because of the transformation in shape of beaker vases between the late-Ming and early-Qing Dynasties (1620-1683), due to changing tastes in this period of dynastic transition. This has captured much scholarly interest. Some scholars, such as Geng Baochang 耿宝昌, Zhu Jun朱军 and Xu Jingjing徐菁菁 have examined the changing shapes of vases in this period, noting that appraisers were required to memorise shapes when inspecting and identifying ceramic vases [7, 8, 9]. Other scholars, such as Soame Jenyns and Margaret Medley, applied a topological method of visual analysis of ceramic vases, leading to a revision in their dating [10, 11]. A combination of quantitative methods and topological techniques were used by Ji Dongge 纪东歌 and Yu Haiyang 于海洋, both of whom were interested in the historical and societal influences over shape design and patterns in the two dynasties [12, 13]. This paper proposes a new, quantitative approach to undertake the study of shapes and forms.

To analyse the dataset, the vases were captured in three dimensions using photogrammetry, from which a 3D model was built. From the mesh of each model, a random sample of vertices of 1000 points was extracted. The vases were roto-translated and centered so that the orientations were standardised across models. These models were compared by relying on metrics that measured the distance between their distribution of points. In this study, an approximation of the Wasserstein metric, known as the Sinkhorn distance, was used. The benefit of the Wasserstein metric for this comparative approach lies in its capacity to synthetise into one 'number' the dissimilarity between two distributions (shapes): the greater the difference, the greater the cost (value) to reposition the points. A pre-existing suite was deployed to implement the algorithm [14, 15]. The Sinkhorn distances are the final output of the analysis. The comparison produced is a series of pairwise distances that can be used to assess the relative closeness or similarity between shapes.

This study has outlined the usefulness of this new computational approach for quantifying changes in East-Asian pottery. The method can be scaled to large datasets of 3D objects scans where changes can be computed automatically, without the need for human intervention. As museums and cultural institutions move to digitise their collections in three dimensions, this approach opens new possibilities for the large-scale study of form across time and geographical locations.

Bibliography

[1] Pala, G. and Costiner, L. (2022). "Tracing Changes in Shape of Historical Artefacts across Time using 3D Scans: A New Computational Approach", *Journal of Open Humanities Data*, forthcoming.

[2] Winckelmann, J. (1764). *Johann Winckelmanns, […] Geschichte der Kunst des Alterthums*. Dresden: In der Waltherischen Hof-Buchhandlung. The English translation is Winckelmann, J. (2006), *The History of the Art of Antiquity.* Los Angeles: Getty Research Institute.

[3] Kubler, G. (1962). *The shape of time: Remarks on the history of things*. New Haven: Yale University Press.

[4] Zhemin, Y., 叶喆民 (2011), *Zhongguo taoci shi* 中国陶瓷史 [*History of Chinese Pottery and Porcelain*]. Beijing: SDX Joint Publishing Company.

[5] Liming, G., L. Hongjie and J. Wilcock (1989). "The Analysis of Ancient Chinese Pottery and Porcelain Shapes: a Study of Classical Profiles from the Yangshao Culture to the Qing Dynasty Using Computerised Profile Data Reduction, Cluster Analysis and Fuzzy Boundary Discrimination", in Rahtz, S. (ed.), *Computer Applications and Quantitative Methods in Archaeology*. CAA89 (BAR International Series 548). Oxford: B.A.R., pp. 362-374.

[6] Liying W., and B. Marwick (2020). "Standardization of ceramic shape: A case study of Iron Age pottery from northeastern Taiwan". *Journal of Archaeological Science: Report,* Vol. 33, pp. 1-11.

[7] Baochang, G. 耿宝昌 (1993). *Ming Qing ciqi jianding* 明清瓷器鉴定 [Ming and Qing Porcelain on Inspection]. Beijing: The Palace Museum.

[8] Jun Z. 朱军 (2002). "*Mingmo Qingchu qinghau huagu jianding* 明末清初青花花觚鉴定 [A late Ming and early Qing dynasty blue and white goblet identification]", *Wenwu Shijie* 文物世界 4, pp. 38-42.

[9] Jingjing, X 徐菁菁 (2017). "*Mingqing cigu yuanliu ji tezheng* 明清瓷觚源流及特征 [Sources and Characteristics of Ming and Qing beaker vases]." *Yishupin* 艺术品, 11, pp. 66-73.

[10] Medley, M. (1987). "The Ming-Qing Transition in Chinese Porcelain", *Arts Asiatiques* 42, pp. 65-76.

[11] Jenyns,S. (1955). "The Wares of the Transitional Period Between the Ming and Ch'ing, 1620-1683", *Archives of the Chinese Art Society of Americas* 9, pp. 20-42.

[12] Dongge, J. (2012). 纪东歌, "*Qingchuqi Jingdezhen Jinian ciqi fenqi yanjiu* 清初期景德镇纪年瓷器分期研究[A staging study of early Qing dynasty Jingdezhen chronological porcelain]", *Zhongguo yishu yanjiu yuan* 中国艺术研究院.

[13] Haiyang, Y. (2012) 于海洋, "*Mingqing guxing ciqi yanjiu* 明清觚形瓷器研究 [A study of Ming and Qing beaker vases]", PhD diss., Jilin Daxue 吉林大学[Jilin University].

[14] Point Cloud Utils (pcu) - A Python library for common, https://github.com/fwilliams/point-cloud-utils.

[15] Cuturi, M. (2013). *Sinkhorn distances: Lightspeed computation of optimal transport. Advances in Neural Information Processing Systems*, 26, pp. 2292-2300.

A Method to Automatically Georeference and Estimate the Coastline Precision of Digital Historical Maps

Pala, Giovanni Maria
giovanni.pala@magd.ox.ac.uk
University of Oxford, United Kingdom

Maps were a key technology in modern navigation yet, until recently, the quantitative study of historical map's content on a large scale has been limited by constraints on access to materials and by computational limitations. Consequently, existing historical studies dealing with cartography have relied on representative examples and curated comparisons, without engaging in formal large-scale investigations (Jupp, 2017; Carlton, 2015; Petto, 2007; Akerman, 2006). The recent flourishing of new digital technologies and materials encourages different approaches (Chiang et al., 2014; Kovarsky, 2011), and has seen a flourishing of methods that explore ways of extracting maps' textual content (see Machines Reading Maps) and achieving semantic segmentation for feature recognition (Petitpierre et al., 2022). In line with recent applications, this contribution presents a new method to register changes in historical maritime cartography by studying coastlines representation. To do so, it introduces a new automatic approach to historical maps georeferencing.

The choice of coastlines avoids issues of tampering: contrary to land boundaries mapping, the coastline was rarely subject to politically driven alterations. It is a very different border. Precision of maps is meant here to be the closeness of the land shapes they represent to current 21 st century representations. Because the geographic coordinate system was, in its general construction, already defined by the late 17th century, it is possible to compare changes of positioning within it across time. Importantly, erosion at high scales is a assumed to be limited.

Rigorously studying the evolution of the quality of coastline mapmaking would contribute to numerous histories related to maritime trade, printing, and the knowledge economy (Kelly and Ó Gráda, 2019; Pascali, 2017; Dowey, 2017). This would also increase our understanding of political connections, especially between distant regions such as Europe and East Asia (Hostetler, 2009).

The measurement of the evolution of coastline and landmass positioning in historical maps requires the solution of three technical problems: georeferencing, segmentation, and assessing the map's precision.

Georeferencing is the process of associating an object (e.g., a map scan) to a system of geographic coordinates. Currently this process is almost invariably done by hand, with the user imputing specific Ground Control Points (GCPs) on the digital image which are associated with the equivalent point of known coordinates on the globe. Existing algorithms can then, with increasing accuracy as the points increase in number, create a geo-referenced raster that is readable by a GIS software (Jenny and Hurni, 2011; Rumsey and Williams, 2002). This process can be very time consuming. The algorithm proposed here solves for the first time this issue by means of a multi-step approach. As a starting point, it uses a state-of-the-art Optical Character Recognition software (OCR) to obtain the longitude and latitude coordinates reported on the boundary of the map. The boundary region is identified with an approach similar to the one used by Ares Oliveira et al. (2018). A Convolutional Neural Network (CNN) modelled after a U-Net encoder-decoder with skip-connections, is then applied on a binary pre-processed version of the map to extract its projection grid. The extracted grid is decomposed in segments via a Probabilistic Hough Transform (Galamhos et al., 1999), and the grid's inclination space is obtained by interpolation of the extracted segments' inclinations. Combining the results, the GCPs are automatically placed at the intersection of the latitude-longitude lines produced from the interpolated grid values. The GCPs' geocoordinate values are the OCR ones.

Once the image is geo-referenced, every image pixel has a latitude-longitude coordinate assigned to it. The pixels, however, need to be categorised (for example, as "land" or "sea"). A semantic segmentation (i.e., by-pixel classification) CNN network has been trained on synthetic data and original map scans to complete the task. The resulting classified and geo-referenced image can be next used to single out the coastlines.

The final step of this process is the measurement of the distance between the segmented map and its 21 st century equivalent. This is the measure of the maps' precision. As a

rule, this study employs a simple error rate built on spatial blocks set at some fixed arc of a degree granularity, but more advanced approaches are considered (e.g., Wasserstein or Sinkhorn distances, see: Villani, 2009; Cuturi, 2013). The error rates obtained for each map, paired with the metadata available for the cartographic objects, create a rich dataset that can be used to model and test hypotheses with other correlates, and improve our capacity to ask questions on the evolution of geographic information, precision, and accuracy.

Bibliography

Akerman, J. R. ed (2006). *Cartographies of Travel and Navigation*. University of Chicago Press.

Ares Oliveira, S., Seguin, B., and Kaplan, F. (2018). dhSegment: A generic deep-learning approach for document segmentation. *Frontiers in Handwriting Recognition (ICFHR), 2018 16th International Conference on*, pp. 7-12

Carlton, G. (2015). *Worldly Consumers: The Demand for Maps in Renaissance Italy*. University of Chicago Press.

Chiang, Y., Leyk, S., and Knoblock, C. A. (2014). A survey of digital map processing techniques. *ACM Computing Surveys (CSUR)*, 47(1): 1-44.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26: 2292-2300.

Dowey, J. (2017). *Mind over matter: access to knowledge and the British Industrial Revolution*. Ph.D. Dissertation, The London School of Economics and Political Science.

Galamhos, C., Matas. J., and Kittler, J. (1999). Progressive probabilistic Hough transform for line detection. *Proceedings: 1999 IEEE computer society conference on computer vision and pattern recognition*, 1: 554-560.

Hostetler, L. (2009). Contending cartographic claims? The Qing empire in Manchu, Chinese, and European maps. In Akerman, J. R. (ed), *The Imperial Map: Cartography and the Mastery of Empire*. University of Chicago Press, pp. 93-132.

Jenny, B., and Hurni, L. (2011). Studying cartographic heritage: Analysis and visualization of geometric distortions. *Computers and Graphics*, 35(2): 402-411.

Jupp, D. L. (2017). Projection, Scale, and Accuracy in the 1721 Kangxi Maps. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 52(3): 215-232.

Kelly, M. and Ó Gráda, C. (2019). Speed under sail during the early industrial revolution (c. 1750–1830). *The Economic History Review*, 72(2): 459-480.

Kovarsky, J. (2011). Searching for early maps: use of online library catalogs. *ACMLA Bulletin*, 138.

Machines Reading Maps, https://www.turing.ac.uk/research/research-projects/machines-reading-maps (accessed 20 April 2022)

Pascali, L. (2017). The wind of change: Maritime technology, trade, and economic development. *American Economic Review*, 107(9): 2821-2854.

Petitpierre, R., Kaplan, F., and di Lenardo, I. (2022). Generic semantic segmentation of historical maps. *CEUR Workshop Proceedings*, http://ceur-ws.org, ISSN 1613-0073.

Petto, C. M. (2007). *When France Was King of Cartography: The Patronage and Production of Maps in Early Modern France*. Lexington books.

Rumsey, D., and Williams, M. (2002). *Historical maps in GIS*.

Villani C. (2009). *Optimal Transport: old and new*. Berlin: Springer, pp. 93-111.

Modeling Chinese Contemporary Calligraphy: the WRITE Dataset

Pasqual, Valentina
valentina.pasqual2@unibo.it
Digital Humanities Advanced Research Centre, University of Bologna

Bisceglia, Marta R.
martarosa.bisceglia@unibo.it
Department of Interpreting and Translation, University of Bologna

Iezzi, Adriana
adriana.iezzi2@unibo.it
Department of Interpreting and Translation, University of Bologna

Merenda, Martina
martina.merenda@unibo.it
Department of Interpreting and Translation, University of Bologna

Tomasi, Francesca
francesca.tomasi@unibo.it
Digital Humanities Advanced Research Centre, University of Bologna

Introduction

Calligraphy has always been the “chief of all Chinese arts” and a central tenet of Chinese civilization. Today, calligraphy is still extremely pervasive in Chinese society. In recent years, many new forms of calligraphy have emerged (in all fields of visual and performing arts) as it has never happened before (Iezzi, 2013). “WRITE” ERC funded project[1] [2] is the first systematic analysis of all these new art forms. Creating the first artworks dataset of these new forms of calligraphy and using a media-based categorization, WRITE will investigate the emergence of these new forms of calligraphy across four collections: (1) “fine and contemporary arts”, (2) decorative and applied arts, (3) performing arts, and (4) graffiti art.

This contribution aims to advance scholarly knowledge and provide new tools for a deeper understanding of Chinese contemporary calligraphy through semantic technologies. WRITE dataset collects, structures and preserves the multifaceted domain of Chinese contemporary calligraphic data composed of artistic, linguistic and socio-political contents. WRITE will examine the innovative ways in which these new forms of calligraphy have responded to, subverted or reinterpreted traditional idioms to define a modern artistic identity that exists comfortably within the global art world while remaining indelibly Chinese (Iezzi, 2015).

State of the art

Linked Open Data has been nowadays a standard in the GLAM domain to increase the value and discoverability of their metadata, fostering reuse and alignment from and to external sources. A growing interest has been expressed towards Chinese asset by outstanding LOD datasets as Pagode Project in Europeana (Bianchi, 2020), Rijks Museum (Dijkshoorn, 2018), British Museum[3]. As aggregators, they provide general metadata excluding a deeper scholar's analysis. In general, many conceptual models have been designed to represent the museal domain, for instance EDM (Charles, 2015), Wikidata (Erxleben et al., 2014). CIDOC CRM (Doerr et al., 2007) and FRBRoo (Bekiari et al., 2015) also broaden the domain to new forms of analysis (e.g. linguistic/textual and artistic analysis).

Method

The WRITE dataset has been designed with the domain experts, by reusing the Wikidata model[4] - a slim yet expressive model, fitting the project requirements and OmekaS templating structure - and extending it for representing WRITE domain information (e.g. calligraphic features as shown in figure 1). An instance of OmekaS[5] (Li, 2020) carrying WRITE data model has been built and shared with domain experts for the collaborative data entry activity. OmekaS has been used as the data registry for WRITE metadata and media.

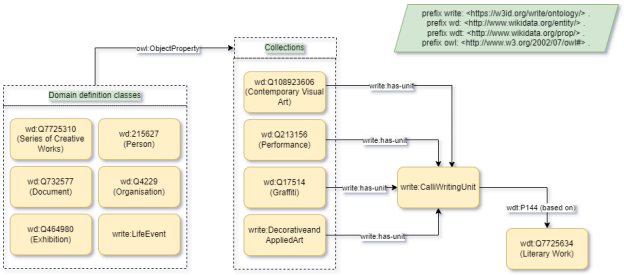


Figure 1:
a summary of the WRITE data model overview

Results

WRITE multilingual (English, Chinese and Pinyin Chinese) dataset stores a set of artworks organized in the four above mentioned collections. Each artwork carries one or more “calli-writing unit(s)”, representing content units with calligraphic features, to store relative artistic and linguistic/textual metadata. Considering the reshaping process of traditional calligraphy in modern production (Iezzi, 2015), such units cannot be categorized as calligraphy as an a-priori assumption. Then, artistic and linguistic/textual metadata over the “calli-writing units” are recorded to classify shared and diverging characteristics with traditional calligraphy as in figure 2.

Class	Calli-Writing Unit
attributed name	CWU The Mountains Are Breaking Up (山摧)
made by material	ink
tool	brush
color	black
background color	white
calligraphic line	Present
script style	en Running script
character form	en Traditional Character
writing system	Chinese characters
transcription or translation	zh 地崩山摧壯士死 [地崩山摧壮士死], 然後天梯石棧相鉤連 [然后天梯石栈相钩连]。 pny Dì bēng shān cuī zhuàngshì sǐ, ránhòu tiān tī shí zhànxiāng gōu lián. en The ground collapsed and the mountain crumbled - warrior heroes died, and only then was this road in the sky linked with a plank road.
significance	Meaningful writing
based on	蜀道難

Figure 2.

OmekaS Item visualization of Calli-Writing Unit with artistic and linguistic metadata of "The Mountains are Breaking Up", Gu Gan (1985).

Additionally, the collections are enhanced through a network of domain and contextual information as exhibitions, artworks series, literary works and agents in their main life events, roles, places, and dates.

WRITE data model documentation[6]and some representative case studies (one artwork for each collection) are available to foster the understanding and reusability of the model. The contribution of the dataset is twofold: first, it supports scholars in investigating the WRITE domain by recording and classifying its data to allow a systematic analysis of its resources to help the domain experts answer their main research questions (e.g. the definition of a spectrum of forms and practices ranging from traditional calligraphy to modern/contemporary artistic expression: can those forms/practices still be defined as calligraphy?) and, hopefully, discover new knowledge; second, it is a challenge in the modelling of different sources from GLAM by categorizing and connecting the visual dimension of iconographic elements (calligraphy as visual artwork) with the respective textual elements (lexical features, text transcription and translation, literary sources of text) to attribute an actual meaning to the new concept of "calli-writing units" through the analysis of the hybridization of multimedia resources.

Bibliography

Bachi, V., Fresa, A. and Veselič, M. (2021). PAGODE – Europeana China. In Ioannides, M., Fink, E., Cantoni, L. and Champion, E. (eds), *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection*. (Lecture Notes in Computer Science). Cham: Springer International Publishing, pp. 265–77 doi: 10.1007/978-3-030-73043-7_22.

Bekiari, C., Doerr, M., Le Bœuf, P. and Riva, P. (2015). *FRBR Object-Oriented Definition and Mapping from FRBRER, FRAD and FRSAD (Version 2.4)*. Final Report International working group on FRBR and CIDOC CRM harmonisation http://www.cidoc-crm.org/frbroo/fm_releases.

Charles, V. and Isaac, A. (2015). Enhancing the Europeana data model (EDM). *EDM WHITE PAPER*.

Dijkshoorn, C., Jongma, L., Aroyo, L., Ossenbruggen, J. van, Schreiber, G., Weele, W. ter and Wielemaker, J. (2018). The Rijksmuseum collection as Linked Data. *Semantic Web*, **9**(2). IOS Press: 221–30 doi: 10.3233/SW-170257.

Doerr, M., Ore, C.-E. and Stead, S. (2007). The CIDOC Conceptual Reference Model - A New Standard for Knowledge Sharing ER2007 Tutorial. , **83**: 51–56 doi: https://doi.org/10.13140/2.1.1420.6400.

Erxleben, F., Günther, M., Krötzsch, M., Mendez, J. and Vrandečić, D. (2014). Introducing Wikidata to the Linked Data Web. In Mika, P., Tudorache, T., Bernstein, A., Welty, C., Knoblock, C., Vrandečić, D., Groth, P., Noy, N., Janowicz, K. and Goble, C. (eds), *The Semantic Web – ISWC 2014*. (Lecture Notes in Computer Science). Cham: Springer International Publishing, pp. 50–65 doi: 10.1007/978-3-319-11964-9_4.

Iezzi, A. (2013). Contemporary Chinese Calligraphy Between Tradition and Innovation. *Journal of Literature and Art Studies*, **3**: 158–79 doi: 10.17265/2159-5836.

Iezzi, A. (2015). What is 'Chinese Modern Calligraphy'? An Exploration of the Critical Debate on Modern Calligraphy in Contemporary China. *Journal of Literature and Art Studies*, **5**: 206–16 doi: 10.17265/2159-5836/2015.03.007.

Li, J. (2020). Omeka Classic vs. Omeka.net. *Emerging Library & Information Perspectives*, **3**(1): 232–36 doi: 10.5206/elip.v3i1.8625.

Notes

1. See https://writecalligraphyproject.eu/
2. The project WRITE - *New Forms of Calligraphy in China: A Contemporary Culture Mirror* is an European Research Council (ERC) Starting Grant funded project based in the Department of Interpreting and Translation of the Alma Mater Studiorum – University of Bologna (GA n. 949645).
3. See https://www.britishmuseum.org/search?search_api_fulltext=chinese+calligraphy
4. See https://www.mediawiki.org/wiki/Wikibase/DataModel
5. OmekaS is a web publishing platform to collaboratively create collections with a shared pool of online resources.
6. See https://write-dataset.github.io/documentation/

Extracting and providing online access to annotated and semantically enriched historical data.The AGODA project

Puren, Marie Anna
marie.puren@epitech.eu

Epitech, MNSHS, France; Centre Jean-Mabillon, Ecole nationale des chartes, France

Vernus, Pierre

pierre.vernus@msh-lse.fr
LARHRA, France; Université Lyon 2, France

Pellet, Aurélien

aurelien.pellet@epitech.eu
Epitech, MNSHS, France

Bourgeois, Nicolas

nicolas.bourgeois@epitech.eu
Epitech, MNSHS, France

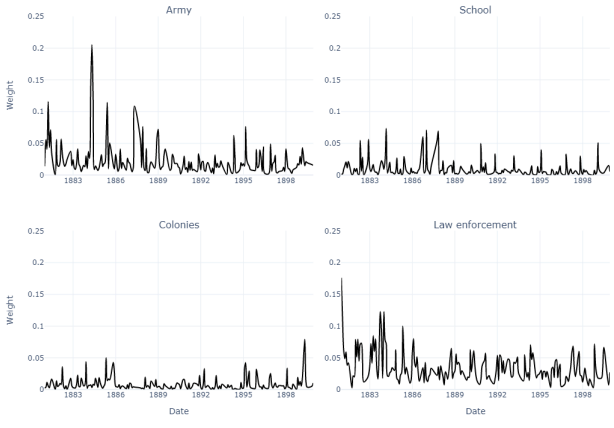
The AGODA project [1] (Puren and Vernus, 2021) is one of five pilot projects supported by the DataLab of the Bibliothèque nationale de France. It aims to create an online platform facilitating the exploration and use of the parliamentary debates of the Chamber of Deputies published in the *Journal officiel* from 1881 to 1940. In the framework of the DataLab, we are working on a test sub-corpus, namely the parliamentary cycle from 1889 to 1893, to test our hypotheses on a smaller dataset.

Over the past sixty years, a great deal of work has been done on parliamentary debates (Chester and Bowring, 1962; Franklin and Norton, 1993). It is indeed a valuable source for historians (Marnot, 2000; Ouellet and Roussel-Beaulieu, 2003; Ihalainen, 2016; Lemercier, 2021), political scientists (Van Dijk, 2010), sociologists (Cheng, 2015) or linguists (de Galembert et al., 2013; Hirst et al., 2014; Rheault et al., 2016). Access to digitised and ocerised debates thus seems to have a positive effect on the number of historical works using these documents (Mela et al., 2022). The same effect can be observed for other disciplines using contemporary debates (Fišer et al., 2018; Fišer et al., 2020). AGODA is thus part of a wider movement to facilitate the use and analysis of parliamentary data, following the example of ParlaClarin (Fišer and Lenardič, 2018) and ParlaMint (Erjavec et al., 2022a; Erjavec et al., 2022b), which propose to produce comparable and multilingual Parliamentary Proceedings Corpora according to the XML-TEI standard. Naomi Truan has also produced a corpus of parliamentary debates encoded in XML-TEI (Truan, 2016; Truan and Romary, 2021). The production of this type of resource facilitates the publication of works exploiting this data to better understand French political discourse (Diwersy et al., 2018; Blaette et al., 2020; Diwersy and Luxardo, 2020).

Between 1881 and 1899, 2596 issues of the *Journal Officiel* were published (50791 JPG images). The debates are also in TXT format but put online without extensive post-correction: the quality of the OCR is not sufficient to provide a satisfactory online browsing experience, and it could have a negative impact on the analyses performed on these texts (van Strien, 2020). Therefore, we chose to ocerise the text, to obtain a better-quality result. We use the PERO OCR (Kodym and Hradiš, 2021; Kohút and Hradiš, 2021; Kišš et al., 2021) based solution developed by the SODUCO project [2]. This tool, still in private alpha version, has been used to prepare the data in (Abadie et al., 2022) that will be accessible via Zenodo.

Ocerised texts are obtained in JSON format; we are developing Python scripts to convert this output into an XML file corresponding to the chosen TEI model. This model is formalised with an adapted XML schema, created using an ODD (Rahtz and Burnard, 2013). We chose to use the ODD created by ParlaClarin (Erjavec and Pančur, 2021) which can be easily adapted to annotate historical parliamentary debates. In the case of France, the rules for transcribing debates were set in the 19th century; thus, the recordings of today's debates are very similar to those produced during the Third Republic. The TEI-encoded corpus will be stored in an eXist-db database, and it will be visualised using the TEI Publisher application, which can transform the source data into HTML web pages. The parliamentary debates will thus be made available to online users as a digital edition and integrated into an application context.

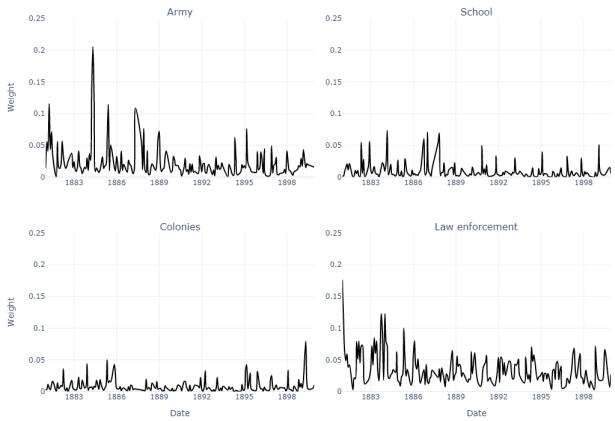
We will also present the first analyses we have carried out on this corpus with "bag-of-words" techniques - these being not too sensitive to the quality of the OCR. We first used topic modelling, an unsupervised learning method that allows us to discover the latent semantic structures of a corpus of texts, without using semantic and lexical resources (Blei et al., 2003). This method is well suited to study parliamentary debates (Bourgeois et al., 2022).



Distribution of four different topics over time

Alternatively, we can use word embeddings to reduce the dimension of the original space from several tens of

thousands of forms to a hundred axes, and then apply classical data science tools such as clustering or correlation analysis on the reduced space (Mikolov et al., 2013). Word embedding has thus shown its interest in the study of parliamentary debates (Rheault and Cochrane, 2020). We used a continuous bag-of-words model for dimension reduction and an unsupervised classification algorithm - in this case DBSCAN - to group words into clusters.



t-SNE projection of the centroïds of the clusters

Bibliography

Abadie, N., Carlinet, E., Chazalon, J., Dumenieu, B. (2022). A Benchmark of Named Entity Recognition Approaches in Historical Documents. Application to 19th Century French Directories. DAS 2022 15th IAPR International Workshop on Document Analysis Systems. La Rochelle, France. May 22-25, 2022.

Blaette, A., Gehlhar, S. and Leonhardt, C. (2020). The Europeanization of Parliamentary Debates on Migration in Austria, France, Germany, and the Netherlands. Proceedings of the Second ParlaCLARIN Workshop. Marseille, France: European Language Resources Association, pp. 66–74 https://aclanthology.org/2020.parlaclarin-1.12 (accessed 21 April 2022).

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation. The Journal of Machine Learning Research, 3: 993–1022.

Cheng, J. E. (2015). Islamophobia, Muslimophobia or racism? Parliamentary discourses on Islam and Muslims in debates on the minaret ban in Switzerland. Discourse & Society, **26**(5). SAGE Publications.

Chester, D. N. and Bowring, N. (1962). Questions in Parliament. Oxford: Clarendon Press.

Diwersy, S., Frontini, F. and Luxardo, G. (2018). The Parliamentary Debates as a Resource for the Textometric Study of the French Political Discourse. Proceedings of the ParlaCLARIN@LREC2018 Workshop. Miyazaki, Japan https://hal.archives-ouvertes.fr/hal-01832649 (accessed 21 April 2022).

Diwersy, S. and Luxardo, G. (2020). Querying a large annotated corpus of parliamentary debates. LREC, ParlaCLARIN Workshop. (Proceedings of the Second ParlaCLARIN Workshop). Marseille, France https://hal.archives-ouvertes.fr/hal-03317717 (accessed 21 April 2022).

Erjavec, T. and Pančur A. (2021) Parla-CLARIN: A TEI Schema for Corpora of Parliamentary Proceedings https://clarin-eric.github.io/parla-clarin/ (accessed 21 April 2022).

Erjavec, T., Pančur A. and Kopp M. (2022a). ParlaMint: Comparable Parliamentary Corpora. GLSL CLARIN ERIC https://github.com/clarin-eric/ParlaMint (accessed 21 April 2022).

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., et al. (2022b). The ParlaMint corpora of parliamentary proceedings. Language Resources and Evaluation doi:10.1007/s10579-021-09574-0.

Fišer, D., Eskevich, M. and Jong, F. de (eds). (2018). Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Paris: European Language Resources Association (ELRA).

Fišer, D., Eskevich, M. and Jong, F. de (eds). (2020). Proceedings of the Second ParlaCLARIN Workshop. Marseille: European Language Resources Association (ELRA).

Fišer, D. and Lenardič, J. (2018). CLARIN resources for parliamentary discourse research. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), pp. 2–7.

Galembert, C. de, Rozenberg, O., Vigour, C. (eds) (2013). Faire parler le parlement: méthodes et enjeux de l'analyse des débats parlementaires pour les sciences sociales. Paris: LGDL-Lextenso.

Hirst, G., Feng, V., Cochrane, C. and Naderi, N. (2014). Argumentation, Ideology, and Issue Framing in Parliamentary Discourse. In Cabrio E, Villata S. and Wyner A. S. (eds), Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing, Forlì-Cesena, Italy, July 21-25, 2014, CEUR-WS.org, http://ceur-ws.org/Vol-1341/paper6.pdf (accessed 26 April 2022).

Ihalainen, P., Ilie, C. and Palonen, K. (2018). Parliament and Parliamentarism: A Comparative History of a European Concept, New York, Oxford: Berghahn.

Ilie, C. (2010). European Parliaments under Scrutiny: Discourse Strategies and Interaction Practices. Amsterdam; Philadelphia: John Benjamins.

Kišš, M., Beneš, K. and Hradiš, M. (2021). AT-ST: Self-training Adaptation Strategy for OCR in

Domains with Limited Transcriptions. In Lladós, J., Lopresti, D., Uchida, S. (eds) Document Analysis and Recognition – ICDAR 2021. ICDAR 2021. Lecture Notes in Computer Science, vol 12824. Cham: Springer, https://doi.org/10.1007/978-3-030-86337-1_31 (accessed 26 April 2022).

Kodym, O. and Hradiš, M. (2021). Page Layout Analysis System for Unconstrained Historic Documents. Document Analysis and Recognition – ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II. Berlin, Heidelberg: Springer-Verlag, pp. 492–506.

Kohút, J. and Hradiš, M. (2021). TS-Net: OCR Trained to Switch Between Text Transcription Styles. In Lladós, J., Lopresti, D., Uchida, S. (eds) Document Analysis and Recognition – ICDAR 2021. ICDAR 2021. Lecture Notes in Computer Science, vol 12824. Cham: Springer, https://doi.org/10.1007/978-3-030-86337-1_32 (accessed 26 April 2022).

La Mela, M., Norén, F., and Hyvönen, E. (2022). Digital parliamentary data in action (DiPaDA 2022), workshop co-located with the 6th Digital Humanities in the Nordic and Baltic countries conference (DHNB 2022), https://dhnb.eu/conferences/dhnb2022/workshops/dipada/ (accessed 26 April 2022).

Lemercier, C. (2021). Un catholique libéral dans le débat parlementaire sur le travail des enfants dans l'industrie (1840). Parlement[s], Revue d'histoire politique, **33**(1): 195–206.

Marnot, B. (2000). Les ingénieurs au Parlement sous la IIIe République. Paris: CNRS Editions.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. ArXiv:1301.3781 [Cs] http://arxiv.org/abs/1301.3781 (accessed 26 April 2022).

Ouellet, J. and Roussel-Beaulieu, F. (2003). Les débats parlementaires au service de l'histoire politique. Bulletin d'histoire politique, **11**(3). Bulletin d'histoire politique: 23–40 doi:10.7202/1060736ar.

Puren, M. and Vernus, P. (2021). AGODA : Analyse sémantique et Graphes relationnels pour l'Ouverture et l'étude des Débats à l'Assemblée nationale. Inauguration Du BnF DataLab. Paris, France https://hal.archives-ouvertes.fr/hal-03382765 (accessed 26 April 2022).

Rahtz, S. and Burnard, L. (2013). Reviewing the TEI ODD system. Proceedings of the 2013 ACM Symposium on Document Engineering. (DocEng '13). New York, NY, USA: Association for Computing Machinery, pp. 193–96.

Rheault, L., Beelen, K., Cochrane, C. and Hirst, G. (2016). Measuring Emotion in Parliamentary Debates with Automated Textual Analysis. PLOS ONE, **11**(12). Public Library of Science: e0168843 doi:10.1371/journal.pone.0168843.

Rheault, L. and Cochrane, C. (2020). Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora. Political Analysis, **28**(1). Cambridge University Press: 112–33 doi:10.1017/pan.2019.26.

Strien, D. A. van, Beelen, K., Ardanuy, M. C., Hosseini, K., McGillivray, B. and Colavizza, G. (2020). Assessing the Impact of OCR Quality on Downstream NLP Tasks. ICAART doi:10.5220/0009169004840496.

Study of parliament group (GB), F., Mark N. and Norton, P. (1993). Parliamentary Questions. Oxford: Clarendon Press.

Truan, N. (2019). Débats parlementaires sur l'Europe à l'Assemblée nationale (2002-2012) [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, v1.1, https://hdl.handle.net/11403/fr-parl/v1.1 (accessed 21 April 2022c).

Truan, N. and Romary, L. (2021). Building, Encoding, and Annotating a Corpus of Parliamentary Debates in XML-TEI: A Cross-Linguistic Account. Journal of the Text Encoding Initiative https://halshs.archives-ouvertes.fr/halshs-03097333 (accessed 21 April 2022).

Notes

1. https://github.com/mpuren/agoda
2. https://soduco.github.io/

Traven between the impostors. Preliminary considerations on an authorship verification case

Rebora, Simone
simone.rebora@univr.it
University of Verona, Italy

Salgaro, Massimo
massimo.salgaro@univr.it
University of Verona, Italy

This paper sets up the groundwork for an authorship verification project dedicated to the German novelist B. Traven, author of novels such as *The Death Ship*(1926) and *The Treasure of the Sierra Madre*(1927), whose real identity is still a mystery. Among the different theories, the most established is the one that sees B. Traven as the pseudonym of Otto Feige, author of a series of political pamphlets for the metal workers' union in Gelsenkirchen,

who then changed his name into Ret Marut (publisher of the anarchist periodical Der Ziegelbrenner), before moving to Mexico and acquiring his final, world-famous pseudonym (Goldwasser, 1993; Hauschild, 2012).

From the point of view of stylometry, that of Feige/Matut/Traven is a typical authorship verification problem, where the goal is not that of attributing an anonymous text to a candidate author, but that of verifying if two (or more) texts were written by the same author. Extensive research is currently dedicated to the subject (see e.g. Kestemont et al., 2020), while an established approach is that of the "impostors" (Koppel and Winter, 2014), recently included also in the stylo package (Eder et al., 2016).

The impostors are intended as a group of writers who cannot be the authors of the texts under examination, but who could function as "distractors" in their attribution. Impostors become useful in authorship verification studies, as they can put into question the attribution of two texts to the same author, even when an alternative candidate author is missing. Main limitation is in their being just surrogates of a possible author, thus never offering a conclusive answer, but only casting doubt on an attribution. The implementation in the stylo package (Eder, 2018) stresses this aspect even further, by providing as result a probability of confirmed attribution which varies slightly at each repetition (due to an element of randomness in the procedure).

To test the efficiency of the impostors in the Traven case study, a corpus was created by collecting essays and articles by Feige, Marut, and Traven, together with texts by Paul Ernst, Erich Mühsam, Robert Musil, and Ludwig Rubiner (see Table 1). This genre limitation was imposed by the fact that Feige wrote just political pamphlets. Ideally, the best impostors would have then been journalists and essayists active at the beginning of the XIX century, but no corpus provided access to this kind of material. The Mannheimer Korpus Historischer Zeitungen und Zeitschriften (Mannheim, 2013) was adopted with the (inevitably forceful) assumption that each journal issue corresponded to one impostor. The Kolimo corpus (Herrmann and Lauer, 2017) offered a more coherent distinction between impostors, but included just fictional and narrative texts (see Table 2).

Author	Source	No. of texts	Total length
Otto Feige	*Metallarbeiter-Zeitung* (1906-1907)	2	2,530 words
Ret Marut	*Der Ziegelbrenner* (1917-1921)	2	11,726 words
B. Traven	*Land des Frühlings* (1928)	2	14,656 words
Paul Ernst	*Theoretischen Schriften* (1905-1931); *Ein Credo* (1935); *Tagebuch eines Dichters* (1934)	3	15,000 words
Erich Mühsam	*Die Psychologie der Erbtante* (1905); *Unpolitische Erinnerungen* (1926)	3	15,000 words
Robert Musil	Journal articles published between 1911 and 1919	3	15,000 words
Ludwig Rubiner	*Der Mensch in der Mitte* (1920)	3	15,000 words

Table 1.
Corpus overview

Corpus	Source	No. of authors	No. of texts
Mannheimer Korpus Historischer Zeitungen und Zeitschriften (zeitungen)	http://repos.ids-mannheim.de/fedora/objects/clarin-ids:mkhz1.00000/datastreams/CMDI/content	NA	652
Kolimo	https://kolimo.uni-goettingen.de/about	6,148	1,176

Table 2.
Imposters corpora

Main idea behind the whole procedure was that of verifying if the impostors could produce high scores when comparing texts written by the same author (e.g. Feige with Feige) and low scores when comparing texts written by different authors (e.g. Traven with Musil). Analyses were performed by combining the 11 distance measures available in the stylo package, 20 different selections of most frequent words (from 50 to 1,000 MFW), 5 culling percentages (0%, 25%, 50%, 75%, 100%), and 8 selections of the number of impostors (from 5 to 40). Computation was repeated 20 times in each condition with the two impostors corpora, thus resulting in a total of 352,000 analyses.

Figure 1 shows a first overview of the results, confirming how in most of the cases impostors work efficiently (with a relatively lower efficiency for Feige).

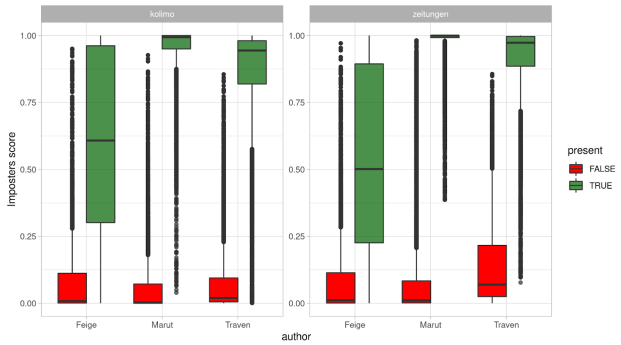


Figure 1.
Testing results

Table 3 confirms how high culling helps overcome the genre differences in Kolimo, while Figure 2 shows how

efficiency increases with MFW selections, but soon reaches a plateau (even at 150 MFW for the Kolimo corpus, 100% culling). Finally, Figure 3 shows how a low number of imposters causes a substantial drop in efficiency (especially for high MFW selections), while more than 10 impostors do not improve the results. Full documentation can be found here: https://github.com/SimoneRebora/Traven_stylometry.

Corpus	Culling	Quality
kolimo	0	0.724
	25	0.725
	50	0.725
	75	0.686
	100	0.741
zeitungen	0	0.735
	25	0.721
	50	0.713
	75	0.642
	100	0.629

Table 3.
Culling results

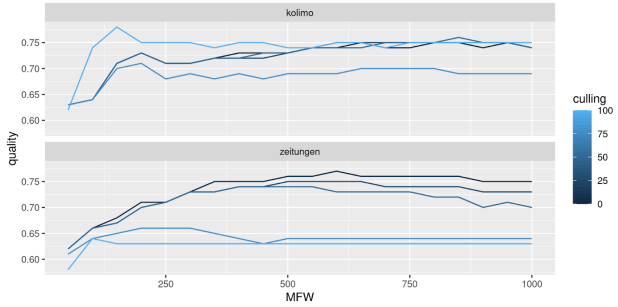


Figure 2.
MFW/culling results

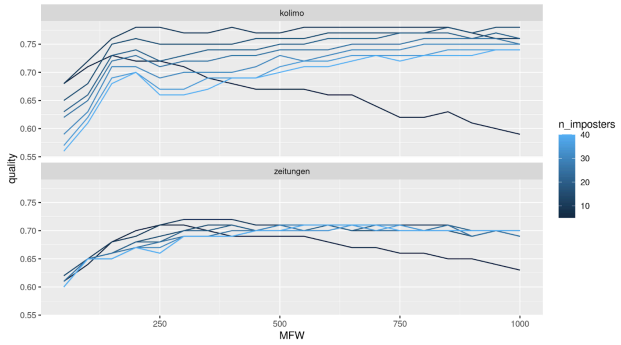


Figure 3.
MFW/impostors number results

The results of this paper will be used for setting up an analysis that needs extensive verification before being put into place, in order to at least minimize the dangers in the application of the impostors method to authorship verification cases like that of B. Traven.

Bibliography

Eder, M.(2018). Authorship verification with the package stylo *Computational Stylistics Group Blog* https://computationalstylistics.github.io/blog/imposters/.

Eder, M., Rybicki, J. and Kestemont, M.(2016). Stylometry with R: A Package for Computational Text Analysis. *The R Journal*, **8**(1): 107–21.

Goldwasser, J.(1993). Ret Marut: The Early B. Traven. *The Germanic Review: Literature, Culture, Theory*, **68**(3): 133–42 doi:10.1080/00168890.1993.9934225.

Hauschild, J.-C.(2012). *B. Traven: Die Unbekannten Jahre*. Zürich : Wien ; New York: Edition Voldemeer ; Springer.

Herrmann, J. B. and Lauer, G.(2017). *KOLIMO. A Corpus of Literary Modernism for Comparative Analysis*. https://kolimo.uni-goettingen.de/about.

Kestemont, M., Manjavacas, E., Markov, I., Bevendorff, J., Wiegmann, M., Stamatatos, E., Potthast, M. and Stein, B.(2020). Overview of the Cross-Domain Authorship Verification Task at PAN 2020. *CLEF 2020 Working Notes*. CEUR, p. 14.

Koppel, M. and Winter, Y.(2014). Determining if two documents are written by the same author: Determining If Two Documents Are Written by the Same Author. *Journal of the Association for Information Science and Technology*, **65**(1): 178–87 doi:10.1002/asi.22954.

Mannheim, I.(2013). *Mannheimer Korpus Historischer Zeitungen Und Zeitschriften*. Mannheim: Institut für Deutsche Sprache Mannheim https://doi.org/10.34644/laudatio-dev-miUsD3MB7CArCQ9C6Cul.

Callimachus: A tool for the study of the formal contents of ancient Greek papyri

Riaño Rufilanchas, Daniel
danielrianno@gmail.com
CSIC, Spain

1. Greek Papyri

The Ancient Greek civilization used numerous materials as support for their texts. Of these, the best preserved are epigraphic texts and the papyri. We can estimate the number of extant papyri to be well over 200,000, of which only a part have been published. They are either literary texts, subliterary texts (private letters, magical texts, etc.) or documentary texts, both public and private (contracts, legal texts, etc.).

Papyri, as we have them now, are very often fragmentary texts, usually poorly preserved. They have transmitted to our days words and expressions that we did not know from literature, and often must be interpreted within a context that can only be conjectured.

Naturally, the chronological extent of the documents testifies to numerous changes in writing practices, and, even more importantly, reflects numerous social changes that took place in the area during these more than a thousand years.

2. Digitization of Greek papyri

A considerable part of all the papyri that have already been scholarly edited have been digitized thanks to several projects, many of them gathered under the aegis of Papyri.info, using as encoding scheme EpiDoc, a subset of the TEI standard adapted to the edition of ancient texts.

Texts in this repository do not identify words separately (other than putting spaces among them), nor do they distinguish normally between words and word fragments, or words and numerals, nor do they lemmatize words.

There are currently more than 81,000 digitized papyri in the Papy.info repository. The site's software allows basic queries within the text of the papyri (without lemmatization) and the metadata that have been incorporated from the Heidelberger Gesamtverzeichnis der Griechischen Papyrusurkunden Ägyptens (HGV) project.

3. Callimachus

Callimachus (https://glg.csic.es/Callimachus/) is a new project currently in development at Spain's CCHS-CSIC, that aims to make available to all users the metadata incorporated in the digital edition of the papyri from Papyri.info, as well as the result of some calculations and estimations about the formal content of the papyri based on such metadata and the data itself. Callimachus processes all digitized papyri and presents the following information using 55 categories.

1. All values and attributes meaningful for the study of Greek papyri contained in the XML markup.

2. It puts this information in relation to the data that about each papyrus has been collected by the HGV.

3. Identifies words and separates them from fragments and numerals.

4. Performs a series of calculations on the textual content of the papyrus, such as the number of words, letters per line, editorial corrections, scribal hands, etc.

4. The *Callimachus number*

The Callimachus number is a decimal number between zero and one that objectively indicates the state of preservation and readability of a document. A Callimachus number 0 indicates no preservation of the contents and 1 indicates that the entire contents of the papyrus have been preserved and are visible without problems. The Callimachus number is based on an algorithm that can be used to determine the state of preservation of any ancient document.

To calculate the Callimachus number we use the XML values and parameters of the document that indicate whether the text has been restored, whether it is fully or partially legible, etc. and perform some kind of calculations to estimate the percentage of missing text. All these calculations result in a matrix that assigns each letter (present or estimated according to the editors' judgment about the extent of the gaps) a value from 0 to 1 according to the criteria in the following table. The sum of all values is divided by the number of letters to obtain the Callimachus number:

5. Some applications of Callimachus

Callimachus can be used in papyrology research as well as in linguistics-related projects. One can use Callimachus to search papyri containing any specific feature, or a combination of features. For example, you can search for papyri containing any specific trait, or combination of traits. The use of Callimachus in combination with Polyphemus will allow to combine lexical information with all the data types of Callimachus.

Polyphemus, a lexical database of the Ancient Greek papyri, and the Madrid Wordlist of Ancient Greek

Riaño Rufilanchas, Daniel
danielrianno@gmail.com
CSIC, Spain

1. Polyphemus, a lexicographic database of Greek papyri

At present time, there is no way to search the corpus of Greek papyri for lemmata, or to search for specific grammatical forms of a word. Much less is there a way to search for examples of a grammatical category. Polyphemus comes to solve these shortcomings, and some more.

For this purpose we have processed all the papyrus texts from PapyInfo (<https://papyri.info/>). This processing is done at the same time as the processing that results in the Callimachus database, which we present at this Congress. I summarize below the procedure by which we obtain our database Polyphemus.

A) First we analyze each line of papyrus and differentiate the actual full words from the gaps or non-textual elements.

B) Then we identify the complete words and separate them from the fragments..

C) We then proceed to lemmatize each of the words, and determine to which part of speech it corresponds, and what is its morphological analysis. All this is done with the help of the Madrid list, which I will discuss below. For text fragments (incomplete words), we try to see if they can be ascribed to a root. We also separate proper nouns from common nouns.

D) Lemma assignment and POS-tagging is performed in two phases. In a first pass we tag the forms with the highest frequency of occurrence. We then go on to label all the remaining forms using the *Madrid Wordlist*.

E) All this information is transferred to a SQL database, and put in relation with the data on the papyri that we have obtained when creating the Callimachus database. In this way, for each lexical form we obtain a lemma, a non-disambiguated morphological analysis, and a translation or gloss. Each of these parameters can be searched in combination with the more than fifty categories available to us thanks to Callimachus, such as date, origin, category, extension, subject, etc.

To date, we have been able to analyze 95% of the complete words, including proper names, which are very numerous.

2. The *Madrid Ancient Greek Word List*

The lemmatization and analysis in Parts Of Speech (POS tagging) is performed by comparing each record in our database with the records of a word list that we have created over the last 3 years, which we have called the Madrid Ancient Greek Wordlist.

Most of the Ancient Greek wordlists are evolutions, simplifications, or improvements from the *Morpheus* list, is a "rule-based morphological analyzer . Our list also starts with Morpheus, but has been enriched with our own treebank (cf. Riaño 2006); the digital version of the *Greek-English Lexicon* of Liddell-Scott-Jones, and Bailly; about 100,000 proper names from *The Lexicon of Greek Personal Names* and the *Trismegistos* repository of papyrological and epigraphic resources. All these data were processed to obtain morphological information: I have generated automatically the Attic and Ionic paradigm for each nominal entry in LSJ and Bailly.

The lemmas are assigned a translation, or rather a gloss, mainly from the *Greek-English Lexicon* of Liddell-Scott-Jones and S.C. Woodhouse "English-Greek dictionary".

3. Polyphemus interface

Polyphemus can be consulted online. It currently contains about 4,600,000 words from Ancient Greek papyri. POS tagging and lemmatization allow the user to query the database for any morphological feature, lemma, or translation. By being able to combine this data with that of the formal content of the papyri provided by the sister database Callimachus, it allows querying the database using more than 80 search criteria.

Since both the original readings and editorial regularizations are preserved, the researcher can use Polyphemus to search for phonetic or morphological features of the papyri. Some searches that can be performed using Polyphemus are the following:

a) Texts containing a Greek word that translates as “poison”, “medicine”, “*praetor*”, “water”, etc.

b) Texts in which any lemma (word) appears, in a specific grammatical form, from Elephantine between the 2nd century BC and 3rd AD.

c) All adjectives in accusative plural; or the optative of verbs in -μι, in all texts.

Bibliography

Bohnet, Bernd and Joakim Nivre 2012 “A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing” *EMNLP-CoNLL,* pp. 1455-1465 [https://aclanthology.org/D12-1133]

Celano, Giuseppe G.A., Gregory Crane and Saeed Majidi 2016 “Part of Speech Tagging for Ancient Greek” *Open Linguistics* 2:393–399 [DOI 10.1515/opli-2016-0020]

Crane, Gregory 1991 “Generating and Parsing Classical Greek” *Literary and Linguistic Computing,* 6:4, pp. 243–245 [https://doi.org/10.1093/llc/6.4.243]

Riaño Rufilanchas, Daniel 2006 *El complemento directo en griego antiguo* en Anejos de Emerita, XLVII. Madrid: CSIC

Close Reading: An Interactive Educational System for Learning How to Read Poetry

Risha, Zak
zak.risha@pitt.edu
School of Computing and Information, University of Pittsburgh

Ma, Rongqian
rom77@pitt.edu
School of Computing and Information, University of Pittsburgh

Introduction

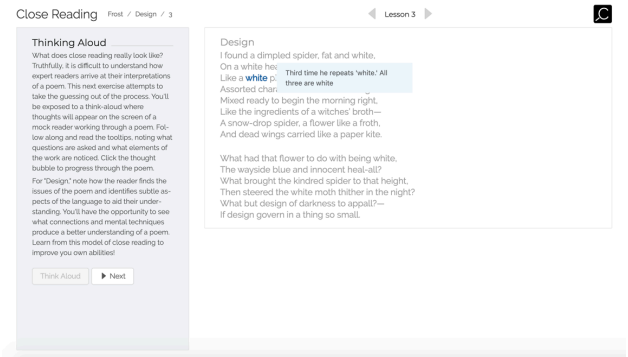
While educational technology research has evolved substantially in the past few decades, there’s evidence that more attention and research has focused on STEM subject areas rather than the humanities (Roll and Wylie, 2016). Additionally, many novel technologies developed by digital humanists have prioritized augmenting research instead of teaching and pedagogy (Fletcher, 2019). Our work seeks to specifically focus on novel educational technology in the humanities by introducing *Close Reading*, a web application for teaching users how to read English poetry. Reading poetry involves additional processes such as identifying and interpreting figurative language and paying close attention to the syntax, sequence of ideas, as well as the structure of the text. *Close Reading* uses a series of interactive pedagogical strategies to cultivate such reading practices in novice readers. Our presentation describes the system interface, activities, and strategies in detail as well as shares current progress and future work for evaluating and piloting the software.

Literature review

Scholars have previously developed related platforms for poetry aimed to augment research and teaching. The *Chinese Text Project* (Sturgeon, 2021) is a dynamic digital library of pre-modern Chinese writing, such as poetry, that builds on the design of crowdsourced collaboration platforms like Wikipedia by adding additional integration with text mining tools to augment research. Using a more interactive approach to engage users, *Jiuge* (Zhipeng et al., 2019) facilitates human machine collaborative generation of Chinese classical poetry that can provide scaffolding to aid beginners in understanding the process of creating poetry. Other systems have more explicitly emphasized the teaching capabilities for poetry, such as *For Better for Verse* (Tucker, 2011) which is a learning system that uses interactive techniques to teach meter (the rhythmic structure of a poem) of English Victorian poems. Likewise, Chen & Lin (2016) designed an educational game for Chinese language poetry education that attempted to provide additional context to poems by simulating aspects of the poets' lives.

In contrast, our system focuses on teaching close reading, a method that emerged from the movement of New Criticism during the 20th century and promotes close and detailed examination of literary texts. It is also a popular pedagogical technique to teach subjects in English literature, and more specifically, poetry (Brooks, 1947; Macey, 2000). Technologies related to the close reading of poetry have sought to support and augment the process for users. *Poemage* (McCurdy et al., 2016) is a visualization tool that allows users to interactively explore the sonic topology of a poem. *Metatation* (Mehta et al., 2017) is a system that augments critic’s annotations as they close read by automatically providing additional information by searching external resources. Our approach focuses more on teaching close reading rather than augmenting it.

System summary

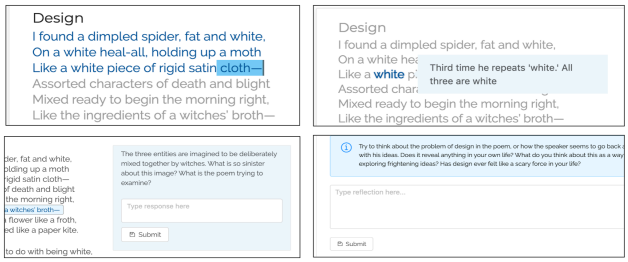


An example of the Close Reading system interface

The *Close Reading* system consists of an interactive interface (Figure 1) where learners advance through four to five activities (Figure 2) relating to a specific poem. These activities are progressive, building on one another in complexity and difficulty.

Pedagogical activities

- **Blocking:** Learners use an interface to block a poem without stanzas into meaningful sections, helping them chunk the text, make note of changes, and summarize the more literal aspects of the poem.
- **Think Aloud:** This activity uses interactive animations to share a mental model of an expert reading and interpreting a poem.
- **Pressure Points:** This activity targets specific parts of the poem relating to the problem/s of the poem, asking the user to respond to a related prompt.
- **Reflection:** Users are asked to apply aspects of the poem to life experiences, reflecting on how the themes and issues of the poem might apply to their own life.



The four activities of *Close Reading*: blocking (top left), think aloud (top right), pressure point (bottom left), and reflection (bottom right).

Current progress and future work

After a prototype was built, we introduced the system to 20 students enrolled in an undergraduate reading poetry course at a university in the Northeastern United States. Our goals were to evaluate these interactive activities and overall usability of the platform. Our preliminary results helped illustrate how the platform could potentially increase students' ability and confidence in interpreting poetry. We report on impressions from students and the challenges of evaluating *Close Reading*'s effectiveness. In the future, we plan to continue refining our evaluation and design methods to better understand how technologies can support the teaching of close reading. While a pedagogically focused project, we believe valuable data will be generated from future user studies and system use for further research in this area.

Bibliography

Brooks, C. (1947). *The Well Wrought Urn: Studies in the Structure of Poetry*. Reynal & Hitchcock.

Chen, H. R. and Lin, Y. S. (2016). An Examination of Digital Game-Based Situated Learning Applied to Chinese Language Poetry Education. *Technology, Pedagogy and Education*, 25(2): 171–86.

Fletcher, C. (2019). Educational Technology and the Humanities: A History of Control. In Gold, M.K. and Klein, L.F. (eds), *Debates in the Digital Humanities 2019*. University of Minnesota Press, pp. 369–81.

Macey, D. (2000). *The Penguin Dictionary of Critical Theory*. Penguin London.

McCurdy, N., Lein, J., Coles, K. and Meyer, M. (2016). Poemage: Visualizing the Sonic Topology of a Poem. *IEEE Transactions on Visualization and Computer Graphics*, **22**(1): 439–48.

Mehta, H., Bradley, A., Hancock, M. and Collins, C. (2017). Metatation: Annotation as Implicit Interaction to Bridge Close and Distant Reading. *ACM Transactions on Computer-Human Interaction*, 24(5): 35:1-35:41.

Roll, I. and Wylie, R. (2016). Evolution and Revolution in Artificial Intelligence in Education. *International Journal of Artificial Intelligence in Education*, 26(2): 582–99.

Sturgeon, D. (2021). Chinese Text Project: A Dynamic Digital Library of Premodern Chinese. *Digital Scholarship in the Humanities*, 36(Supplement_1): i101–12.

Tucker, H. F. (2011). Poetic Data and the News from Poems: A For Better for Verse Memoir. *Victorian Poetry*, 49(2): 267–81.

Zhipeng, G., Yi, X., Sun, M., Li, W., Yang, C., Liang, J., Chen, H., Zhang, Y. and Li, R. (2019). Jiuge: A Human-Machine Collaborative Chinese Classical Poetry Generation System. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Florence, Italy: Association for Computational Linguistics, pp. 25–30.

Asia News in the 16th and 17th Centuries: Catalogue and Digital Library

Rojo-Mejuto, Natalia
natalia.rojo@udc.es
Universidade da Coruña, Spain

Garrobo-Peral, Manuel
manuel.garrobo@udc.es

Università degli studi di Verona, Italy; Universidade da Coruña, Spain



The aim of this presentation is to show the improvements in cataloguing and editing Asia news published during the sixteenth and seventeenth centuries within the framework of the digital project CBDRS, Catálogo y Biblioteca Digital de Relaciones de Sucesos (https://www.bidiso.es/CBDRS). This project, part of BIDISO (Biblioteca Digital Siglo de Oro) and active since 1996, offers open access to Early Modern literature, in particular, pre-periodical press, printed works known as *avvisi*, *nuova* or *relazioni* (Italy), *Zeitung* (Germany), *tijding* (Low Countries), *advis* (France), *relación de sucesos* (Spain), *news* (England), or *nova* (Portugal) (cf. Arblaster et al., 2016).

Keeping in mind both pan-European communication during the Early Modern period and pan-Asian diversity, this presentation focuses on the news related to Japan, China, India, Korea, Taiwan, Vietnam, and Cambodia, among other regions. The diffusion of Asian news sheets, especially during the seventeenth century is attested in their several editions and reprints across Europe. Rapidly translated from one language to another, these documents are a valuable source to know the deep dimensions of cultural exchange throughout Asia and Europe (Rojo-Mejuto, 2021a, 2021b).

For that reason, since 2020, the localization of news pamphlets on Asia has grown exponentially in the CBDRS project. At present, there are more than 200 editions, mainly related to Kirishitan presence in Japan, with one or more surviving copies catalogued and localized in libraries and other centers across the world (Rojo-Mejuto, 2019). CBDRS provides bibliographical and typo-bibliographical descriptions of the works, library's signatures, classification by main genre (political and religious events, ceremonies and festivities, travels, and extraordinary events), a filter for languages (Spanish, Portuguese, Italian, Latin, French, German), georeferentiation of the places mentioned in the texts, as well as descriptions of the illustrations included.

The relational database is in continuous process of technologies and models adaptations to expand and to improve the data as well as the digital objects. It is possible to refine the search by the name and variants of the persons involved in the creation of the pamphlets (authors, translators, editors, printers) and to manage the uncertainty of the date of the event and of edition with the attributes "estimated" and "deduced" (Pena-Sueiro and Saavedra-Places, 2019). Likewise, searching by place of edition or location of event allows to find editions in the map (https://www.bidiso.es/CBDRS/lugares).



We will improve exponentially our possibilities of analysis, manipulation, and reception with the implementation of XML-TEI editions, especially because this standard offers us two main advantages (Fernández-Travieso and Garrobo-Peral, 2021):

A) Registration of all the characteristics of the news that we want to label in order to make compatible our main goals with the awareness that its use is not linked to a specific digital setting that may provoke loss of information.

B) Manageability when feeding the preservation of our work, the migration or reuse of our data and the inter-operability with other projects, thus treating XML-TEI as an input-output format.

In sum, the presentation addresses one of the main problems of working on sixteenth and seventeenth news of Asia, solving the dispersion of surviving copies in libraries, and providing access in one site to all these primary sources. Besides bibliographic matters, the work emphasizes the vitality of news networks between Asia and Europe, as well as other important press locations, such as Manila or Mexico.

CBDRS, as other projects on early modern news (cf. Folch, 2006; Euronews Project), hopes to continue facilitating an easy and fruitful access to the knowledge treasured in these texts to scholars and society. Thus,

another objective of this presentation is to showcase the possibilities of the Digital Humanities project CBDRS as a starting point for research works on Anthropology, Culture, History, Linguistics, and Press, besides the social interest of knowing the birth of journalism, that is, the birth of a way of shaping information, thinking, and international communication.

Bibliography

Arblaster, P. et al. (2016). The lexicons of Early Modern News. In Raymond, J. and Moxham, N. (eds.), *News Networks in Early Modern Europe*. Leiden/Boston: Brill, pp. 64-101. https://doi.org/10.1163/9789004277199.

EURONEWS Project. https://www.euronewsproject.org/.

Fernández-Travieso, C. and Garrobo-Peral, M. (2021). Creación de un corpus digital de textos de relaciones de sucesos: un modelado de datos XML-TEI para acoger diferentes tipos de edición. In Bazzaco, S. (coord.), Congreso Internacional Humanidades Digitales y Estudios Literarios Hispánicos, Verona 22-23 June. Università degli Studi di Verona.

Folch, D. (2006). La China en España. Elaboración de un corpus digitalizado de documentos españoles sobre China de 1555 a 1900. https://www.upf.edu/asia/projectes/che/principal.htm.

Pena-Sueiro, N. and Saavedra-Places, Á. (2019). Obsolescencia y resilencia en Humanidades digitales. El caso de la Biblioteca Digital de Relaciones de Sucesos. *Artnodes*, 23: 79-88. http://dx.doi.org/10.7238/a.v0i23.3243.

Rojo-Mejuto, N. (2019). Las relaciones de sucesos españolas sobre Japón en los siglos XVI y XVII. In Torres, L., Tropé, H. and Espejo-Surós, J. (coords.), IX Coloquio de la Sociedad Internacional para el Estudio de las Relaciones de Sucesos (SIERS). Rennes 19-21 September. Université Rennes 2.

Rojo-Mejuto, N. (2021a). Japón escondido: cosas muy notables en las cartas de las Indias. In Andrés-Renales, G. and Peñasco-González, S. (eds.), *Buenas noticias: relaciones de sucesos en los siglos XVI-XVIII*. Pesaro: Metauro, pp. 49-64.

Rojo-Mejuto, N. (2021b). La literatura informativa sobre Japón en 1621. In Borrego, M. (coord.), Seminario Internacional El mundo en 1621: avisos, relaciones de sucesos y conexiones culturales. Besançon 16-17 September 2021. Université de Franche-Compté.

Mining and Modeling Spaces and Places for Literary History as Linked Open Data

Röttgermann, Julia
roettger@uni-trier.de
Trier University, Germany

Hinzmann, Maria
hinzmannm@uni-trier.de
Trier University, Germany

Dietz, Katharina
dietz@uni-trier.de
Trier University, Germany

Gebhard, Henning
gebhard@uni-trier.de
Trier University, Germany

Klee, Anne
klee@uni-trier.de
Trier University, Germany

Konstanciak, Johanna
konstanciak@uni-trier.de
Trier University, Germany

Schöch, Christof
schoech@uni-trier.de
Trier University, Germany

Steffes, Moritz
steffesm@uni-trier.de
Trier University, Germany

Introduction

In literary history and historiography, places and spaces play an important role – not least in the context of the 'spatial turn' (Lafon, 1997; Piatti et al., 2009; Dennerlein, 2009; Weber, 2014). In literary works, narrative locations are particularly relevant, but places of publication as well as further spatial dimensions can also be taken into account (Curran 2018; Burrows et al. 2016). Our contribution presents how we obtained spatial statements from three

different information sources and combined them in a knowledge network based on the Linked Open Data (LOD) paradigm (Berners-Lee, 2006; Hooland und Verbough, 2014; Hitzler, 2021).

Project context

The aim of the project Mining and Modeling Text is to establish an information network for the humanities built from various sources.[1] This aim is closely linked to the finding that, considering the steadily growing digital cultural heritage, the acquisition of knowledge from large amounts of text and data can no longer be handled by individuals. In representing knowledge as LOD, we see untapped potential that we are exploring in the current project phase on the French Enlightenment novel (Delon/ Malandain, 1996; Mylne, 1981).

Creating spatial statements

Information on spatial statements relevant to our domain is extracted from three different types of sources: 1. bibliographic metadata, 2. primary sources, 3. scholarly publications. After focusing on thematic statements in an earlier phase of our project (Schöch et. al., 2022; Röttgermann et al., 2022), we are currently addressing spatial statements.

Mining

Bibliographic metadata: For our domain, the Bibliographie du genre romanesque français 1751-1800 (BGRF, Mylne et al., 1977) is central, as it defines the population of about 2000 French Enlightenment novels. The BGRF has been extensively analysed and modeled (Lüschow 2020) and contains rich metadata (including places of publication, narrative locations, narrative form, characters, themes, style).

Primary sources: The Collection of Eighteenth-Century French Novels (Röttgermann, 2021) is analysed via SpaCy's (Honnibal und Montani, 2017) named entity recognition and reconciliation pipeline supported by OpenRefine (Huynh, [2012] 2010). Our pipeline requires human intervention (Hinzmann et al., 2022) concerning the challenges of ambiguity, fictionality and historicity (Heuser et al., 2016; Jockers, 2016; Nielsen, 2016).

All		Column 1	bgrf	LOC1	Wikidata Label fr	URL	Quantity LOC1
1.		Abbes_Voyage	58.5	Palais	palais	https://www.wikidata.org/wiki/Q16560	3
2.		Anonym_Suzon	83.9	Couvent	couvent	https://www.wikidata.org/wiki/Q1128397	21
3.		Anonyme_Zoloe	00.37				
4.		Arnaud_Epoux	83.15	Paris	Paris	https://www.wikidata.org/wiki/Q90	27
5.		Arnaud_Matinees	99.43				
6.		Arnaud_Sentiment	70.21	Nancy	Nancy	https://www.wikidata.org/wiki/Q40898	111
7.		Barthelemy_Voyage	88.27	Grèce	Grèce	https://www.wikidata.org/wiki/Q41	440

Figure 1:
Reconciliation of narrative locations with OpenRefine

Modeling

Our approach relies on importing both the text strings as found in the information source (green) and the abstract spatial items (blue) into our Wikibase instance. Combined with the fact that we reference all statements uniquely via the "stated in" property (orange), this ensures a high degree of verifiability of our data (see fig. 2).

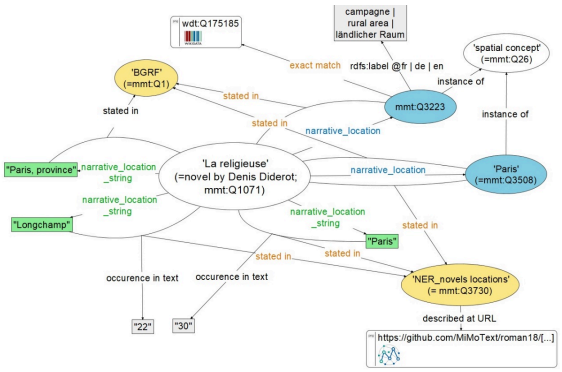


Figure 2:
'Narrative locations' of Diderots La religieuse from NER and BGRF data

Our spatial vocabulary was built up incrementally and provides items (=spatial concepts) for 7 properties, of which 5 are currently mapped with Wikidata (see fig. 3). [2]

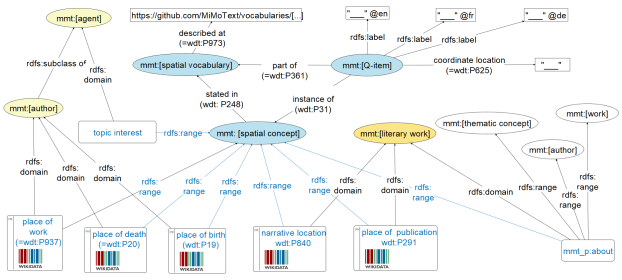


Figure 3:
Ontology of 'spatial statements' within the domain of literary history/historiography

Infrastructure

For the provision of data, we follow Open Science principles, such as the publication of FAIR data in open access as well as the use of open source software – in particular Wikibase (see fig.4). [3] We created a custom bot using the Python library Pywikibot to import and update the RDF triples into our Wikibase instance from TSV files. [4]

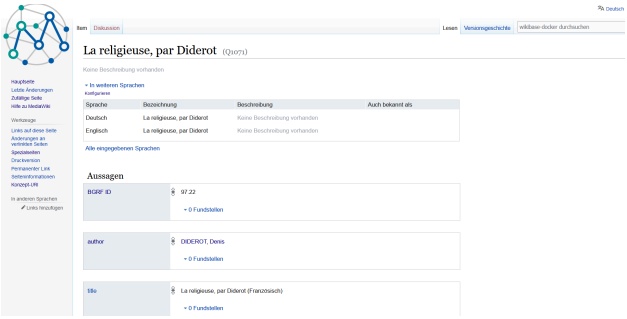
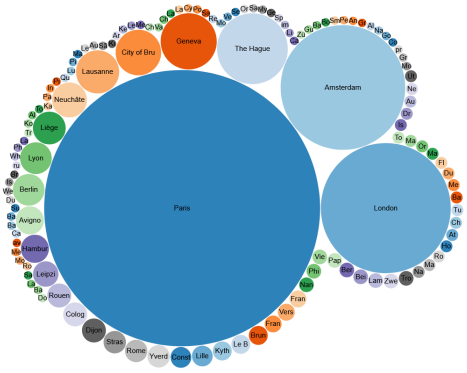


Fig. 4:
Local Wikibase instance

Spatial Querying

Having all the spatial triples stored in our Wikibase allows us to query and visualize them using the DockerWikibaseQueryService interface. We can gain an overview of the entire set of metadata (see fig. 5, query 1), see places of publication appearing and disappearing over time (see fig. 6, query 2) or explore narrative locations linked to specific thematic concepts such as ‘miracle’ (see fig. 7, query 3). Via 'federated queries' (see fig. 8, query 4), information from other knowledge bases (here Wikidata) can be used.


Paris
London
Amsterdam
The Hague
Geneva
City of Bru
Lausanne
Neuchâte
Liège
Lyon
Berlin
Avigno
Leipzi
Rouen
Colog
Dijon
Stras
Rome
Yverd
Lille
Kyiv
Le B
Brun
Vers
Fran
Pau
Phi
Ve
Ma
Ro
Ch
Tu
Me
Fl
Gr
To
Na
Au
Ba
Ca
Ge
Ha
Ka
Ki
Ar
Ut
Ni
Ze
Gr-Sha
Ha
Ro
Ma
Ge
Ka
Ch
Ar
Le-Ha
Ba
Ge
Qu
Ba
Ka
Ki
Tr
Ha
Wh
Ha
Ha
We
Da
Ha
Ca
Ha
Ma
Ro
Ha
Da
Ha
Ha
Ha

Appendix

Query 1: Items (novels) and their place of publication (fig. 5)

```
#defaultView:BubbleChart
SELECT ?topLabel (count(*) as ?count)
WHERE {
 ?item wdt:P8 ?top . #P8 = 'place of publication'
 ?top rdfs:label ?topLabel .
 filter(lang(?topLabel) = "en")
}
GROUP BY ?topLabel
ORDER BY desc(?count)
```

Query 2: Places of publication over time (fig. 6)

```
#defaultView:BarChart

Select ?year ?placelabel(count(*) as ?countyear)
   WHERE{
   ?item wdt:P8 ?place.
   ?place rdfs:label ?placelabel .
   ?item wdt:P7 ?date .
   FILTER(lang(?placelabel) = "fr")
   BIND(str(year(?date)) as ?year)
   SERVICE wikibase:label {bd:serviceParam wikibase:language "[AUTO_LANGUAGE]","fr" .}
  }

GROUP BY ?year ?placelabel
having (?countyear > 1)
#ORDER BY asc(?year) desc(?countyear)
```

Query 3: Narrative location of novels with theme "miracle" (fig. 7)

```
#defaultView:TreeMap
SELECT
   ?value ?valueLabel ?work ?workLabel
WHERE
{
 ?work wdt:P52 ?value; #exploring narrative location of works
    wdt:P25 wd:Q2970 #works with miracle as theme
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
}

```

Query 4: Places of publication with geocoordinate location via federated query (fig. 8)

```
#defaultView:Map{"markercluster":"true"}
PREFIX wid: <http://www.wikidata.org/entity/> #wikidata wd
PREFIX widt: <http://www.wikidata.org/prop/direct/> #wikidata wdt

Select DISTINCT ?item ?itemLabel ?loc ?locLabel ?wikidataEntityLink ?coordinateLocation ?theme ?themeLabel
{
  ?item wdt:P8 ?loc. # loc = publication place
  ?loc wdt:P30 ?WikiLink.

  BIND(IRI(REPLACE( STR(?WikiLink),"https://www.wikidata.org/wiki","http://www.wikidata.org/entity" )) AS ?wikidataEntityLink).
  #Federated Query -> Wikidata
  SERVICE <https://query.wikidata.org/sparql> {
    ?wikidataEntityLink
                  widt:P625 ?coordinateLocation
  }

  SERVICE wikibase:label {
    bd:serviceParam wikibase:language "en" .
  }
}

```

Bibliography

Berners-Lee, T. (2006). *Linked Data – Design Issues*. https://www.w3.org/DesignIssues/LinkedData.html.

Burrows, S. et al. (2016). Mapping Print, Connecting Cultures. *Library & Information History*, 32(4), pp. 259–71. 10.1080/17583489.2016.1220781.

Curran, M. (2018). *The French Book Trade in Enlightenment Europe I: Selling Enlightenment*. London: Bloomsbury Publishing.

Delon, M. and Malandain, P. (1996). *Littérature française du XVIIIe siècle*. Paris: Presses universitaires de France. https://gallica.bnf.fr/ark:/12148/bpt6k48060529.

Dennerlein, K. (2009). *Narratologie Des Raumes*. Berlin, New York: De Gruyter.

Heuser, R., Algee-Hewitt, M. and Lockhart, A. (2016). Mapping the Emotions of London in Fiction, 1700–1900: A Crowdsourcing Experiment. In *Literary Mapping in the Digital Age*. London: Routledge.

Hitzler, P. (2021). A Review of the Semantic Web Field. *Commun. ACM*. 10.1145/3397512.

Honnibal, M. and Montani, I. (2017). *SpaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing*.

Hooland, S. van. and Verborgh, R. (2014). *Linked Data for Libraries, Archives and Museums: How to Clean, Link and Publish Your Metadata*. Facet Publishing.

Huynh, D. (2010). *OpenRefine*. OpenRefine. https://github.com/OpenRefine/OpenRefine.

Jockers, M. L. (2016). The Ancient World in Nineteenth-Century Fiction; or, Correlating Theme, Geography, and Sentiment in the Nineteenth Century Literary Imagination . *Digital Humanities Quarterly*, 010(2).

Lafon, H. (1997). *Espaces Romanesques Du XVIIIe Siècle, 1670-1820: De Madame de Villedieu à Nodier*. 1. éd. Paris: Presses Universitaires de France.

Martin, A., Mylne, V. and Frautschi, R. L. (1977). *Bibliographie du genre romanesque français, 1751-1800*. London: Mansell.

Mylne, V. (1981). *The Eighteenth-Century French Novel: Techniques of Illusion*. 2nd Edition. Cambridge, New York: Cambridge University Press.

Nielsen, F. A. (2016). Literature, Geolocation and Wikidata. In *Wiki@ICWSM*. Proceedings of the International AAAI Conference on Web and Social Media. Cologne, pp. 61–4. https://ojs.aaai.org/index.php/ICWSM/article/view/14833.

Röttgermann, J. (ed.) (2021). Collection de romans français du dix-huitième siècle (1750-1800) / Eighteenth-Century French Novels (1750-1800) [dataset]. *Release v0.2.0*. 10.5281/zenodo.5040855.

Schöch, C. et al. (2022). Smart Modelling for Digital Literary History. *IJHAC: International Journal of Humanities and Arts Computing [Special Issue on Linked Open Data]*, 16(1), pp. 78–93. https://doi.org/10.3366/ijhac.2022.0278.

Suber, P. (2012). *Open Access*. Cambridge, Mass: The MIT Press.

Weber, A.-K. (2014). *Mapping Literature: Spatial Data Modelling and Automated Cartographic Visualisation of Fictional Spaces*. Zurich: ETH Zurich. 10.3929/ETHZ-A-010106067.

Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data*, 3(1), p. 160018. 10.1038/sdata.2016.18.

Notes

1. See https://mimotext.uni-trier.de/en/.
2. See for the vocabulary: https://github.com/MiMoText/vocabularies/blob/main/spatial_vocabulary.tsv.
3. Generally, see Suber 2012 and Wilkinson et al. 2016; related to the project, see Röttgermann and Schöch 2020 and Schöch, 2021. We expect the Wikibase instance to become publicly available in mid-2022: https://www.mimotext.uni-trier.de/en
4. See https://github.com/MiMoText/Wikibase-Bot.

The Silences in Archives: A Case Study of Annual Reports from Archives at the National Centre for Biological Sciences, India

Roy, Dibyadyuti
dibyadyutir@gmail.com
Indian Institute of Technology (IIT) Jodhpur, India

Taparia, Kanishtha
taparia.1@iitj.ac.in
Indian Institute of Technology (IIT) Jodhpur, India

Srinivasan, Venkat
venkats@ncbs.res.in
Archive at the National Centre for Biological Sciences (NCBS)

Contemporary Digital Humanities Scholarship has placed considerable emphasis on the power of archives that has its etymological root in the Greek word "Arkhe"—simultaneously connoting commencement and commandment—as not only sites of preservation but also as opportune sites of exclusion. Institutional archives in particular, through their exclusion of certain records and narratives, create epistemic gaps that enable dominant narratives while concurrently silencing minority voices, events, and subjectivities. In particular for scientific and research institutions, institutional archives and repositories provide crucial evidence of the challenges overcome in the pursuit of scientific knowledge as well as the subtle social changes precipitated by the growth of science and technology in local contexts. Contextually in India, colonial legacies combined with the vagaries of postcolonial statecraft ensure that many of the leading scientific and research institutes in India do not have any institutional archives. A notable outlier is the Archives at The National Center for Biological Sciences (NCBS), which is a collecting archive chronicling the growth of the NCBS along with being a site to document the history of contemporary biology in India. The NCBS was established in 1992 and is a part of the Tata Institute of Fundamental Research, Mumbai (TIFR) which is a National Centre of the Government of India, under the umbrella of the Department of Atomic Energy.

While considerable importance is given in archival and DH scholarship to first person accounts of scientists and similar dominant stakeholders, there is hardly any scholarship that looks at the annual reports of scientific institutions as knowledge systems that generate new questions or inaugurate new models of inquiry. Further existing scholarship has drawn attention toward the underrepresentation of female scientists in archival spaces: but have rarely parsed how institutionalized gender biases mediate such silences. In addressing these lacunae, this project will narrate the results, complications and future directions of an exploratory study which analyses the digitized annual reports for a period of 25 years (1992-2017)

from the Archives of the National Center for Biological Sciences (NCBS).

There is sparse research that evaluates the socio-cultural undertones of an organization evidenced through its annual reports (Jagolinzer, 2021). Although the key purpose of annual reports is to provide public-domain disclosure of the operations and financial activities of the organization over the year, they are also opportune sites to understand institutional biases and silences. In focusing on the annual reports of NCBS over a 25-year period this project draws attention to this underappreciated genre that is often described as the most common medium for communicating strategies and plans, past performance, and future expectations of an organization (Srinivasan and Marques, 2017). In particular for research institutions, annual reports not only provide periodic research updates but also documents the subterranean and hegemonic policies, framing research frameworks and decisions. The analytical lens employed in this study employs the intersecting domains of Digital Humanities, Gender Studies, and STS (Science and Technology Studies). We employ CTDA (Critical Techno-cultural discourse analysis) as the guiding methodology to identify and recover the gendered institutional silences, biases, and underappreciated female subjectivities which often remain elided within normative research on and about scientific institutions. The CTDA triad of *technology*, *technology in practice*, and *socio-cultural beliefs* is operationalized with digital archives, the act of archiving and hegemonic masculinity representing the three facets respectively. With the use of this methodology attention is drawn towards the ways women scientists are perceived in institutional archival spaces.

By articulating significance and interpretation of archives in digital spaces, we examine the discourse of gender representation in digitally archived objects such as these annual reports. Further, to qualitatively assess the issue of women's representation in science, we operationalize techno-feminism (combining STS and Feminism) asserting that the technological advancements and social circumstances are not mutually exclusive. With an emphasis to understand gendered institutional silences, we also supplement our analysis with interviews of women scientists from NCBS. Our intervention combines a longitudinal computational analysis of digitized annual reports from the NCBS archive over a 25-year period with qualitative interviews to exemplify the potential of Critical Digital Humanities: to not only shed new light on the gendered and unexplored facets of scientific research in Global South spaces but also develop unique methodological approaches for insightful methods of knowledge creation in DH.

Bibliography

Brock, A. (2016). Critical technocultural discourse analysis: *Sage Publication*, *20*(3), 1012–1030.

Jagolinzer, A. (2021). Annual reports should inform society – not only those with a financial interest.

Jasanoff, S. (2004). The Idiom of Co-Production. In *States of Knowledge: The Co-Production of Science and Social Order* (pp. 1–12). Taylor & Francis.

Srinivasan, P., & Marques, A. (2017). *Narrative analysis of annual reports: A study of communication efficiency* (No. 486).

Wajcman, J. (2010). Feminist theories of technology. *Cambridge Journal of Economics*, *34*(1), 143–152.

The benefits of increasing the digital availability of Alsatian theater

Ruiz Fabo, Pablo
ruizfabo@unistra.fr
Université de Strasbourg, France

Werner, Carole
wernerc@unistra.fr
Université de Strasbourg, France

Bernhard, Delphine
dbernhard@unistra.fr
Université de Strasbourg, France

Extensive efforts have been made to create digital corpora for major European dramatic traditions. A recent corpus hub is DraCor (Fischer & Börner, 2019), offering programmatic access to theater collections in 11 languages. [1]

The creation of such corpora was accompanied by computational literary research on their traditions. Focusing on German and French drama (immediately relevant for our work), one may name projects like DLINA (Fischer et al., 2017), [2] QuaDramA and Q:TRACK (Reiter, 2020) [3] or *Emotions in Drama* (cf. Schmidt et al., 2021), for German; for French, studies like Schöch's (2017), or works in the *Revue Historiographique du Théâtre*'s special issue on Digital Humanities and theater (Galleron, 2017). [4]

Alsatian refers to Germanic varieties from Alsace (Eastern France), the main oral communication language

in the area for over 14 centuries, until their decline vs. French in the last third of the 20th century (Huck et al., 2007: 11–17). Despite Alsatian's mainly oral status, it has a rich theatrical tradition, where comedic and popular genres predominate (Gall, 1974). Unlike major European traditions, digital resources for Alsatian theater are scarce, which precludes related computational literary research. To help make such research feasible, our project [5] is building a large TEI-encoded corpus of Alsatian theater, annotated with rich metadata, and making this corpus publicly available in formats appropriate for research and for the non-specialist community, covering mostly the 1870-1940 period. Our proposal discusses the relevance of developing such materials, in terms of opportunities for (computational) literary studies (CLS) and natural language processing (NLP), besides their potential to promote linguistic diversity through the digital medium.

Our corpus both complements and challenges existing knowledge of Alsatian theater. Earlier studies on the corpus period (Cerf, 1972, 1975; Huck, 1998, 2005; Hülsen, 2003) examined plots, plays' social groups, and dialect representation in character speech, based on corpora with under 40 plays. Annotating the characters' socio-professional groups in 109 plays (Ruiz Fabo and Werner, 2021) revealed the importance of artisans in plays set in small towns, contrary to claims in earlier studies, which presented agricultural professions as almost exclusive in such settings (Cerf, 1972: 340). The corpus also has potential for comparisons with the two hegemonic dramatic traditions that surround Alsatian theater: French and German. Hülsen (2003: 98 ff.) described the influence of German popular comedy in Alsatian theater, based on thirteen major Alsatian plays around 1900. Our corpus could be used for a larger-sample comparison, covering a longer period, with German and French plays from the same timeframe. A concrete first aspect to compare could be the *dramatis personae*, following our work on character annotation, cited above.

Digital literary collections for under-resourced languages also benefit NLP. Current NLP models like contextual embeddings (e.g. Devlin et al., 2018) require large amounts of text, yet unavailable for languages like Alsatian. However, the NLP community is interested in improving the viability of these technologies for low-resource languages, and in developing corpora in such languages (cf. Bernhard et al., 2019). Increasing the availability of Alsatian electronic text contributes to this effort and is a test-bed for new technology developments.

Towards our goals, we selected a varied corpus mixing major and lesser-known works. We performed optical character recognition (OCR) on National Library digitizations, [6] creating 25 TEI plays (300,000 tokens). OCR was also manually corrected for 28 further plays, awaiting TEI conversion. Bibliographical metadata were transcribed for 359 plays, and the settings and *dramatis personae* for 257 plays (2,253 characters). We also annotated characters' social variables when possible (gender, professional group, estimated social class, age). Synopses were written for 38 plays. TEI materials are publicly available [7] and accepted at the DraCor platform. Publication with a DOI on open repository Nakala was also carried out. [8]

A corpus exploration interface exposes all data. [9] Reiter et al. (2017) warn that a graphical user interface does not always help literary research. However, our project intends to promote public interest in Alsatian beyond research, and we believe that a web interface is a suitable means. Our interface gives access to all plays with corrected OCR output, TEI-encoded or not, searching by author, author gender, dates or publisher. It also allows filtering the corpus based on the presence of characters with different social attributes (professional group, social class, gender, see Figure 1). We believe that such navigation workflows are an attractive way for the public to engage with the corpus and with the social picture it conveys, and will assess this hypothesis with a user sample. Character co-occurrence is also relevant for research on plot and for sociolinguistic studies on character speech representation according to social variables.

Regarding limitations, male authors predominate, even if we sought variety. There are only five female authors, with 14 plays. Deeper research in non-digitized collections should counter this bias. Note also that sexism and racism are present in the corpus, and, even if the interface targets the general public, it does not at this point provide means that would help non-specialists be aware of these biases and critically question such social representations. This is important future work.

In sum, we present a digital corpus of plays in Alsatian varieties with rich metadata, including character annotations for social variables. TEI encoding is ongoing; where not available, corrected OCR output is provided. This material helps CLS research on Alsatian theater, impossible before given lack of a corpus, and has shown potential to challenge existing knowledge on the tradition. The texts also have relevance for NLP research on low-resource languages. To promote interest in the corpus' language varieties, a corpus exploration interface opens up these resources to the community.



Character cooccurrences in cast list per professional group. Top: Plays by female authors. Bottom: Plays by male authors. Proportion of female characters with profession in cast list is 45.71% with female authors and 11.30% with male authors.

Acknowledgements

We thank further project members, Dominique Huck and Pascale Erhart, for advice on corpus selection and other project aspects. We also thank all interns who contributed to corpus or interface development: Nathanaël Beiner, Andrew Briand, Lena Camillone, Hoda Chouaib, Audrey Deck, Barbara Hoff, Valentine Jung, Salomé Klein, Audrey Li-Thiao-Te, Kévin Michoud and Vedisha Toory.

Bibliography

Bernhard, D., Bras, M., Erhart, P., Ligozat, A.-L. and Vergez-Couret, M. (2019). Language Technologies for Regional Languages of France: The RESTAURE Project. *International Conference Language Technologies for All (LT4All): Enabling Linguistic Diversity and Multilingualism Worldwide*. Paris, France https://hal.archives-ouvertes.fr/hal-02418928 (accessed 14 January 2020).

Cerf, E. (1972). Le théâtre alsacien de Strasbourg, miroir d'une société (1898-1939). *Saisons d'Alsace*(43).

Cerf, E. (1975). Les contes merveilleux du théâtre alsacien de Strasbourg. *Revue des sciences sociales de la France de l'Est*, **4**: 3–30.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]* http://arxiv.org/abs/1810.04805 (accessed 6 December 2020).

Fischer, F. and Börner, I. (2019). Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. *Digital Humanities 2019*. Utrecht, p. 5.

Fischer, F., Göbel, M., Kampkaspar, D. and Kittel, C. (2017). Network Dynamics, Plot Analysis: Approaching the Progressive Structuration of Literary Texts. *Digital Humanities 2017*. Montreal, pp. 437–41.

Gall, J.-M. (1974). *Le théâtre populaire alsacien au XIXe siècle*. (Recherches et Documents XIX). Strasbourg: Istra.

Galleron, I. (ed). (2017). *Revue d'Historiographie Du Théâtre, Numéro 4 : Études Théâtrales et Humanités Numériques*.

Huck, D. (1998). D'r Herr Maire (1898) de Gustave Stoskopf. Entre ethnologie et littérature : les Alsaciens en auto-représentation. *Recherches Germaniques*, **28**: 163–90.

Huck, D. (2005). Le «Théâtre Alsacien de Strasbourg» et la production dramaturgique de ses fondateurs (1898-1914). *Culture et Histoire Des Spectacles En Alsace et En Lorraine : De l'annexion à La Décentralisation (1871-1946)*. Peter Lang, pp. 198–222.

Huck, D., Bothorel-Witz, A. and Geiger-Jallet, A. (2007). L'Alsace et ses langues. Eléments de description d'une situation sociolinguistique en zone frontalière. In Abel, A., Stuflesser, M. and Voltmer, L. (eds), *Aspects of Multilingualism in European Border Regions: Insights and Views from Alsace, Eastern Macedonia and Thrace, the Lublin Voivodeship and South Tyrol*. Bozen/Bolzano: EURAC Research (Europäische Akademie / Accademia Europea / European Academy), pp. 13–101 http://ala.u-strasbg.fr/documents/Publication%20-%20L%27Alsace%20et%20ses%20langues.pdf.

Hülsen, B. von (2003). *Szenenwechsel im Elsass: Theater und Gesellschaft in Strassburg zwischen Deutschland und Frankreich 1890-1944*. Leipziger Universitätsverlag.

Reiter, N. (2020). Möglichkeiten Quantitativer Dramenanalyse. *Comparatio. Zeitschrift für Vergleichende Literaturwissenschaft*, **12**(2): 39–52 doi: 10.1007/978-3-476-04516-4_16. https://comp.winter-verlag.de/article/COMP/2020/2/7 (accessed 7 December 2021).

Reiter, N., Kuhn, J. and Willand, M. (2017). To GUI or not to GUI?. *INFORMATIK 2017*. Gesellschaft für Informatik https://dl.gi.de/bitstream/handle/20.500.12116/3880/B14-8.pdf?sequence=1&isAllowed=y (accessed 13 November 2019).

Ruiz Fabo, P. and Werner, C. (2021). Exploration du théâtre alsacien à travers ses listes de personnages pendant la période 1870-1940. *Humanistica 2021*. Rennes

(online): Zenodo doi: 10.5281/zenodo.4762733. https://doi.org/10.5281/zenodo.4762732 (accessed 8 December 2021).

Schmidt, T., Dennerlein, K. and Wolff, C. (2021). Emotion Classification in German Plays with Transformer-based Language Models Pretrained on Historical and Contemporary Language. *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Punta Cana, Dominican Republic (online): Association for Computational Linguistics, pp. 67–79 https://aclanthology.org/2021.latechclfl-1.8 (accessed 7 December 2021).

Schöch, C. (2017). Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. *Digital Humanities Quarterly*, **011**(2).

Notes

1. https://dracor.org/doc/creditslists corpora and languages. Our first 25 TEI-encoded Alsatian plays were also accepted in DraCor.
2. https://dlina.github.io/
3. https://quadrama.github.io/
4. *Journal for Theater Historiography*. Special issue: https://sht.asso.fr/revue/etudes-theatrales-et-humanites-numeriques/
5. MeThAL, "Towards a macroanalysis of theater in Alsatian" (https://methal.pages.unistra.fr/en.html)
6. Numistral portal, BNU Strasbourg: https://numistral.fr/fr/theatre-alsacien
7. https://git.unistra.fr/methal/methal-sources
8. Nakala set https://nakala.fr/collection/10.34847/nkl.feb4r8j9 contains our data
9. https://methal.eu/ui/

Modeling Music for Musicologists: A Linked Open Data Approach

Saccomano, Mark
mark.saccomano@uni-paderborn.de
Paderborn University, Center for Music, Edition, Media (ZenMEM)

Shibata, Elisabete
shibata@beethoven.de
Beethoven-Haus Bonn, Research Centre "Beethoven-Archiv"

Lewis, David
david.lewis@oerc.ox.ac.uk
University of Oxford e-Research Centre

Hankinson, Andrew
andrew.hankinson@rism.digital
Répertoire International des Sources Musicales (RISM) Digital Center

Page, Kevin
kevin.page@oerc.ox.ac.uk
University of Oxford e-Research Centre

Modeling work in the digital humanities has traditionally focused on written texts; music, however, requires data models that can capture the varied, overlapping layers that are characteristic of its structure. Most encoding models for notated music—i.e., those presenting a precise representation of what can be performed by a musician—provide good coverage of the layers on the immediate music 'surface,' like measures, staves, and notes. Other, more analytical and less apparent structures are often not as well addressed. This includes formal objects such as musical themes and motifs, as well as properties that often lack explicit means for symbolic notation, such as texture and timbre. Our paper describes a model that includes a component to specifically address these types of musical structure, using different arrangements of the same musical work as examples.

Musical figures require descriptions that include both the beginning and end points that mark their extent, as well as a specification of the individual parameters lying at different structural layers that comprise that figure. For example, early arrangements of Beethoven's Eighth Symphony contain a replacement of the novel triple forte dynamic marking with a simple fortissimo, or double forte, at the moment of the first movement's recapitulation. We can point to the respective measures and state that one has a different dynamic marking than the other, but this does not include our identification of the measures as different expressions of the same music-theoretical structure: a recapitulation. Although digital annotations can reference passages in multiple works, these references do not make sense unless they are linked together as separate expressions of the same analytic object. In order to present such a comparison, we need to posit a distinct, abstract class to model it. Such parallelism is entailed in our common understanding of what an arrangement is, though as Flanders and Jannidis (2021) note, the modeling already inherent in such a term is 'invisible through [its] very familiarity.' Before beginning work on a data model,

therefore, we are obliged to examine what it is we are looking for when examining 'versions.'

Our model builds on Linked Data principles demonstrated in projects using the Music Encoding and Linked Data (MELD) framework and consists of three modules: one for identifying resources, one for scholarly annotation, and, at the core, a framework for categorizing, labeling, and comparing user-selected structural features along with their formal analogues in different arrangements. After considering other standards, we have aligned our framework with the Functional Requirements for Bibliographic Records (FRBR), adding specialized subclasses of the standard FRBR entities for use in music comparison. Our model's music component introduces a class at the FRBR Expression level so that targeted commentary can be attached not simply to a contiguous block of music, but also, for example, to symbols that indicate the manner in which an accompanying melodic figure is to be played, or to the repetition and variation of a certain theme within a single piece.

The data model has further been designed to accommodate source materials in different formats, including standardized methods to refer to segments (e.g., Media Fragments, EMA, and IIIF): by collecting individuated classes at the manifestation level, a specific portion of music can be treated independently of its realization in different media. This way, an annotation can target a semantically meaningful musical selection across file formats, rather than a set of resource-specific IDs.

A prototype using the model has been successfully developed for a digital musicology study of 19th-century arrangements of orchestral works: https://github.com/DomesticBeethoven/bith-annotator. This application lets a user view scores encoded as MEI files side-by-side. They can then select a group of measures from two different versions of the same work and mark them as corresponding excerpts of the same passage of music. These excerpts are then saved as a single object—a parallel passage—ready for scholarly annotation.

Targeting musicological objects in an encoding involves more than capturing the symbols present in a particular region of a printed score. It entails the selection of parameters that are constitutive of the object, yet cannot be specified in advance. By introducing a class that incorporates these features into a single object with a musicologically meaningful label, this data model allows such abstract structures to be compared to one another in multiple versions and multiple files. In addition, because the model is compatible with Linked Data formats, these objects can be reused in future research, thus providing digital musicology projects with the potential to have a greater, longer-lasting contribution to scholarship in the field.

Funding: This research was undertaken by the project *'Beethoven in the House: Digital Studies of Domestic Music Arrangements*,' and supported by a UK-Germany funding initiative: in the UK by the Arts and Humanities Research Council (AHRC) [project number AH/T01279X/1], and in Germany funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) [project number 429039809].

Bibliography

Flanders J. and Jannidis, F. (2019). Data Modeling in a Digital Humanities Context: An Introduction. In Flanders, J. and Jannidis, F. (eds.), *The Shape of Data in the Digital Humanities: Modeling Texts and Text-based* Resources. New York: Routledge, 2019, pp. 26–96.

International Image Interoperability Framework (IIIF). (2020). API Specifications. https://iiif.io/api / (accessed April 28, 2022).

Lewis, D., Page, K. and Dreyfus, L. (2021). Narratives and Exploration in a Musicology App: Supporting Scholarly Argument with the Lohengrin TimeMachine. In *8th International Conference on Digital Libraries for Musicology* (DLfM '21). New York: Association for Computing Machinery, pp. 50–58.

Music Encoding and Linked Data (MELD). (2021). Overview. https://meld.web.ox.ac.uk/ (accessed April 28, 2021).

Music Encoding Initiative (MEI). (2019). https://music-encoding.org/ (accessed April 28, 2022).

Viglianti, R. (2015). Enhancing Music Notation Addressability (EMA). https://music-addressability.github.io/ema/ (accessed April 28, 2022).

Weigl, D., et al. (2021). Notes on the Music: A Social Data Infrastructure for Music Annotation. In *8th International Conference on Digital Libraries for Musicology* (DLfM '21). New York: Association for Computing Machinery, pp. 23–31.

W3C Media Fragments Working Group. (2012). Media Fragments URI 1.0. https://www.w3.org/TR/media-frags/ (accessed April 28, 2022).

Detecting Latent Textual Bias with Topic Modeling and Sentiment Analysis

Sanders, Ashley
asandersgarcia@ucla.edu
UCLA, United States of America

Bias detection is an emerging area of research for digital humanists, computational linguists, and information studies scholars, alike, who point to biases inherent in our algorithms, software, tools, and platforms, but we are only just beginning to examine how computational methods could be used to interrogate our primary textual sources (Noble, 2018; Al-Sarraj and Lubbad, 2018; Chen et al., 2020). This project seeks to develop a method for bias detection that can be used at the outset of a study with little initial knowledge of the corpus, requires little pre-processing and is both beginner-friendly and language-agnostic. Word2Vec and similarity measures allow us to compare a test corpus against a comparison corpus of biased or neutral terms. This works especially well with contemporary texts, such as online news articles in English, but it becomes an increasingly difficult task with historical or non-English language sources to find appropriate comparative corpora (Patankar and Bose, 2017). Building classifiers to identify bias with feature extraction, support vector machine learning algorithms, decision trees, and naïve Bayes approaches work well but require a deep understanding of the corpus and are not accessible to those who are new to computation. (Al-Sarraj and Lubbad, 2018; Leavy, 2019; Manzini et al., 2019) Therefore, with the aforementioned aims in mind, I chose to use the latent Dirichlet allocation (LDA) algorithm for topic modeling to study a set of three chronicles covering the 300 years of Ottoman Algerian history, written in French by two nineteenth-century French scholars and one twentieth-century Algerian scholar (Vayssettes, 1867; Mercier, 1903; Gaïd, 1978). [1]

At just 138 documents and approximately 183,000 words, the corpus is much smaller than one normally uses for topic modeling, but its manageable size made it a promising test case for this approach. The scalar feature of the LDA algorithm was particularly interesting to examine, as the topics of each larger, more detailed model neatly nested under the model with the next most topics, creating a hierarchy. [2] Nesting models of 4-, 7-, 11-, and 20-topics provided a detailed summary of the corpus, with the 4-topic model serving as a general overview, and the 20-topic model offering a glimpse into the richness of the region's history with topics related to some of the dominant themes of the corpus, such as "governance and succession," but also much more granular themes, including "illness, death, burial, and remembrance," and the "roles of women." Pairing topic models at different scales (4-, 7-, 11-, and 20-topics) with sentiment analysis of the topic models at 11- and 20-topics, as well as targeted close reading, guided by topics of interest and using the concordance method to identify passages with key terms of interest uncovered the stories of lesser known actors, including women, Jews, Spaniards, and the councilmen of provincial governors, as well as biases inherent in the writing of their histories (Ghasiya and Okamura, 2021).

The anti-Arab and/or anti-Turkish sentiments one might expect to observe were absent, but a latent anti-Semitic sentiment appeared in the more granular topic models that, despite my careful reading of the texts, had escaped my notice. The resulting model aids the scholar in weaving the disparate threads of these individuals' lives into the tapestry of the region's history, and the method may well be applied to other corpora, topics, languages, and time periods to reveal hidden biases, especially in larger collections of documents that would be impossible for a single scholar to read. I am currently testing this bias detection method with additional corpora and this presentation will briefly report the results of these trials along with the original study.

Bibliography

Al-Sarraj, W. F. and Lubbad, H. M. (2018). Bias Detection of Palestinian/Israeli Conflict in Western Media: A Sentiment Analysis Experimental Study. *2018 International Conference on Promising Electronic Technologies (ICPET)*. pp. 98–103 doi:10.1109/ICPET.2018.00024.

Chen, W.-F., Al-Khatib, K., Stein, B. and Wachsmuth, H. (2020). Detecting Media Bias in News Articles using Gaussian Bias Distributions. *ArXiv:2010.10649 [Cs]* http://arxiv.org/abs/2010.10649 (accessed 12 May 2021).

Ghasiya, P. and Okamura, K. (2021). Understanding the Middle East through the eyes of Japan's Newspapers: A topic modelling and sentiment analysis approach. *Digital Scholarship in the Humanities* (fqab019) doi:10.1093/llc/fqab019. https://doi.org/10.1093/llc/fqab019 (accessed 19 June 2021).

Guldi, J. and Williams, B. (2018). Synthesis and Large-Scale Textual Corpora: A Nested Topic Model of Britain's Debates over Landed Property in the Nineteenth Century. *Current Research in Digital History*, **1** doi:https://doi.org/10.31835/crdh.2018.01. https://crdh.rrchnm.org/essays/v01-01-synthesis-and-large-scale-textual-corpora/ (accessed 11 January 2021).

Leavy, S. (2019). Uncovering gender bias in newspaper coverage of Irish politicians using machine learning. *Digital Scholarship in the Humanities*, **34** (1): 48–63 doi:10.1093/llc/fqy005.

Manzini, T., Lim, Y. C., Tsvetkov, Y. and Black, A. W. (2019). Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. *ArXiv:1904.04047 [Cs, Stat]* http://arxiv.org/abs/1904.04047 (accessed 12 May 2021).

Nielsen, R. (2019). Quantitative Text Analysis in Arabic Workshop Cairo University http://www.mit.edu/~rnielsen/arabic_text_slides.pdf (accessed 19 May 2021).

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press (accessed 19 May 2021).

Patankar, A. A. and Bose, J. (2017). Bias Discovery in News Articles Using Word Vectors. *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. pp. 785–88 doi:10.1109/ICMLA.2017.00-62.

Notes

1. For more details on the authors, see the introductory chapter and/or the README on my GitHub repository for this project, available at https://github.com/AshleySanders/OttomanAlgeria/ . It should also be noted that I have not included the one additional Algerian chronicle of these governors, written in Arabic, because the LDA algorithm separates topics by language first and then by significant collocations, which would not help me identify common themes that cut across three centuries, individual biographies, and sources/authors. For those interested in topic modeling a collection of Arabic documents, see Richard Nielsen's introduction using R: Richard Nielsen, "Quantitative Text Analysis in Arabic," (Workshop, Cairo University, April 4, 2019), accessed May 6, 2021, http://www.mit.edu/~rnielsen/arabic_text_slides.pdf.
2. The resulting nested or hierarchical model visualizes major and minor themes in the authors' perceptions and presentations of Ottoman gubernatorial histories. Jo Guldi and Benjamin Williams applied a similar approach to British Parliamentary discourse to reveal previously invisible connections between speeches and political tactics, but, as this chapter shows, the method is also useful for text summarization and bias detection.(Guldi and Williams, 2018)

Representing scholarly statements in ontologies for data management: The case of musicology

Sanfilippo, Emilio M.
emilio.sanfilippo@cnr.it
Laboratory for Applied Ontology ISTC-CNR, Italy

Freedman, Richard
rfreedma@haverford.edu
Haverford College, USA

Music analysts need to document **statements** about subjects relevant in their area of expertise. And since researchers frequently assume different and even contradictory perspectives on the same subject, the representation of both the **work** and the intellectual responsibility for claims made about it is crucial. The use of Semantic Web (SW) technologies has made data more accessible and discoverable. But only recently has anyone attempted to model interpretive claims (Cristofaro et al., 2021). The *Ontology for Analytic Claims in Music* (OMAC) is a SW ontology (under development) to fill this gap, proposing innovative ways of modeling both musical works and the interpretive arguments about them. For space limits, we cannot document the entire ontology here. We will limit to the introduction of some aspects; readers can refer to the Web repository for more information (https://github.com/emiliosanfilippo/OMAC).

At the current state, OMAC consists of two modules, the Musical Work module for musical works, and the Analytic Claim module for scholarly arguments. The latter can take the form of assertions about, e.g., **authorship**, **chronology**, and **similarity**, all of which figure in qualitative arguments made by analysts and critics. OMAC can be extended with further elements relevant from a musical perspective, including performances. To maximize the reuse of SW resources, we adopted elements from DBpedia, Dublin Core, etc. We have not reused ontologies based on FRBR (Bekiari et al., 2017) like the Music Ontology (Raimond et al., 2007) because of the ambiguous manner in which FRBR treats works (Sanfilippo, 2021).

The Musical Work module in OMAC models the **authorial** structure of works, namely, the division by the composer of a work into (sub-)sections (Fig. 1). A Renaissance Mass, for instance, consists of customary five sections (*Kyrie*, *Gloria*, *Credo*, *Sanctus*, and *Agnus Dei*), some of which are in turn divided into subsections. No claim about this subdivision is necessary, as they are dimensions of the authorial text. Once modeled in OMAC, it is possible to reason over data, e.g., automatically deducing the structure of a whole work.

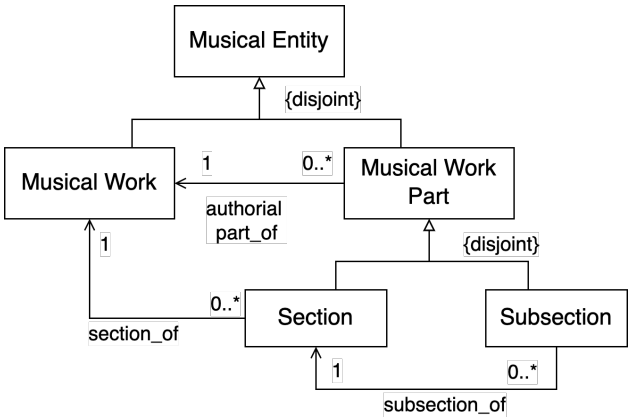


Figure 1
Authorial structure of musical entities (partial view)

Analytic claims could relate to an entire work or some of its parts. Two different scholars, for instance, might each argue that a work was by a different composer or that it was created in one year or another. Scholars also might want to make a claim about a single pattern in a piece, fragments that could never be called authorial. For instance, participants in the research *Citations: The Renaissance Imitation Mass* (http://crimproject.org/) are interested in identifying small passages in pairs of works, showing how one composer borrowed from (and transformed) the work of another. In the context of this analytic project, scholars make specific claims about the connections between what they call a **model** and its **derivative**. But such assertions are in principle a claim about **similarity**, and as such should be made discoverable as instances of this more general musical principle. Indeed, scholars might want to make many different kinds of similarity claims, which might concern borrowing, or quotation, or simply shared style. In OMAC such claims can be modeled in a logical way that gives critical assertions a declared (and thus computable) structure, as shown in Figure 2.

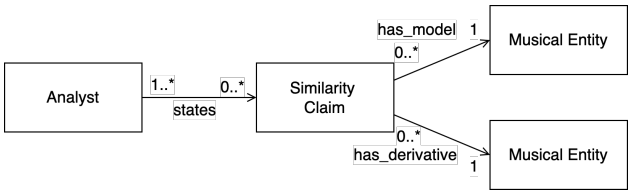


Figure 2
Similarity claims in OMAC (partial view)

OMAC takes its place in the broader context of recent discussion of how best to model critical assertions (and not simply works) in a digital environment. A proposal by Masolo et al. (2018), for instance, argues that scholarly observations are **epistemological states**, i.e., the classification of an entity under a property as the result of an analytic process. For instance, the similarity relation between two musical entities is an observation made by an analyst but it does not represent something that is necessarily true. In addition, nothing prevents an analyst from reviewing or discarding their claim, or other analysts in formulating different and conflicting claims about the same phenomena.

Future research will enrich the structure of OMAC (including an explicit treatment of events) and contribute to the development of a digital space to share musicological data. In addition, as a lesson learnt from OMAC, the representation of claims can be generalized and tuned to other areas in the humanities in such a way to make the ontology broader and reusable across projects.

Bibliography

Bekiari, C., Doerr, M., Le Boeuf, P., and Riva, P. (2017). Definition of FRBRoo: A conceptual model for bibliographic information in object-oriented formalism

Cristofaro, S., Sanfilippo, E.M., Sichera, P., and Spampinato, D. (2021). Towards the representation of claims in ontologies for the digital humanities. In *Proc. of the Int. Joint Workshop on Semantic Web and Ontology Design for Cultural Heritage (SWODCH)*, CEUR vol. 2949

Masolo, C., Botti Benevides, A., and Porello, D. (2018). The interplay between models and observations. *Applied Ontology*, *13*(1), 41-71

Raimond, Y., Abdallah, S. A., Sandler, M. B., and Giasson, F. (2007). The Music Ontology. In *ISMIR* (Vol. 2007, p. 8th)

Sanfilippo, E. M. (2021). Ontologies for information entities: State of the art and open challenges. *Applied ontology*, vol. 16, no. 2, pp. 111-135, 2021

Visualising Women's Lives : A Feminist Approach to Distant Reading

Schreibman, Susan
susan.schreibman@gmail.com
Maastricht University, Netherlands, The

Barget, Monika
m.barget@maastrichtuniversity.nl
Maastricht University, Netherlands, The

This presentation describes ongoing research which adopts a distant reading approach to recovering the lives

of women in early 20th century Ireland. Our case studies are based on a comparatively small and diverse sub-corpus of letters made public by the *Letters 1916-1923* project, a dataset of some 6000 letters from public and private sources in Ireland and beyond. The first case study concerns women who were active in supporting the war effort during the First World War from 1916-1918 and the second revolves around a central male figure, Charlie Daly, a Republican combatant in the Irish Civil War (1922-1923)

This research is a conscious act of recovery that combines the analysis of anecdotal thoughts and feelings (common in traditional scholarship that uses letters as a primary source) with different methods of distant reading (Moretti 2013; Drucker 2017a;). By including sources beyond the *Letters 1916-1923* collection (e.g. newspapers and reports), we get a broader picture of the networks in which the women operated, how gender norms changed, culturally, politically, and socially, from the years of the Great War to the post-war period. Moreover, we contribute to a growing feminist discourse in Irish historiography by reinstating the contributions of female non-combatants in two wars that were pivotal in shaping the modern Irish state (Walsh, 2020, Cullen 2017; Pašeta 2017). This research takes an iterative approach to reading which we call *reading at the middle distance*: eg. looking for patterns by distant reading the data to find points of interest for close reading. This approach to text analysis overlaps with the concept of "scalable reading" promoted by digital historians who seek "a critical engagement with data-driven historical scholarship" and combine "different 'modes' of reading." (Fickers & Clavert, 2021) Our feminist approach also requires more than one method of quantitative text analysis: apart from descriptive statistics based on our metadata categories and tags, we apply topic modelling and mapping to trace women's activities and their networks.

In creating the visualisations that support the research process, we take a critical visualisation approach, being aware that the visualisations we produce are the results of an inherently remediated and interpretative process (Dork et al 2013; d'Ignazio & Klein 2016) in which lacuna and missing information is difficult if not impossible to represent (eg – visualising the absence of data is inherently impossible as what is not available remains an unknown). Contextualising our visualisations in our public GITHUB repository (Barget and Schreibman 2020) is, therefore, an important step. The data tables behind our visualisations make the (ambiguous) sources of our findings transparent and highlight cases in which incomplete information in the letters was supplemented with information from other primary sources or secondary works. Thus, the resultant data sets and the visualisation formats we choose become agents in meaning-making, embedding implicit knowledge into the design process itself (Drucker 2017b).

This short paper will highlight the connection that recent theoretical developments in data feminism and interactive, explorative *scalable reading* have with our perception of *middle distance reading*. While our work on women's stories in the Letters 1916-1923 collection began prior to the publication of *Data Feminism (2020)*, the principles that D'Ignazio and Klen outline [1] prove a useful touching stone from which to examine biases in data and reflect our own subjectivity as authors and researchers (Leurs 2017). The concept of *Data Feminism* stressed our central claim that using methods of quantitative text analysis comparatively and in new contexts, rather than to reify bias in the records, provides a creative and disruptive perspective on women's experience, surfaces alternative narratives of historical periods, and re-evaluates power dynamics and hierarchies of influence.

Bibliography

Barget, Monika, and Schreibman, S. (2020) 'Women's Agency and Networks in Ireland (1915-1923)'. Github Pages. FeministDH. https://monikabarget.github.io/FeministDH/.

Cullen, C (2017). War Work on the Home Front: The Central Sphagnum Depot for Ireland at the Royal College of Science for Ireland, 1915–1919'. *Medicine, Health and Irish Experiences of Conflict, 1914–45*, ed. David Durnin and Ian Miller (Manchester: Manchester University Press, 155–70.

D'Ignazio, C., & Klein, L. F. (2016). Feminist data visualization. Workshop on Visualization for the Digital Humanities (VIS4DH), Baltimore. IEEE.

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT press.

Dörk, M., Feng, P., Collins, C., & Carpendale, S. (2013). Critical InfoVis: exploring the politics of visualization. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems* (pp. 2189-2198).

Drucker, J. (2017a). Why Distant Reading Isn't. *PMLA/Publications of the Modern Language Association of America*, *132*(3), 628-635.

Drucker, J. (2017b). Information visualization and/as enunciation. *Journal of Documentation*.

Fickers, A., & Clavert, F. (2021). On pyramids, prisms, and scalable reading. *Journal of Digital History*, *jdh001*. https://journalofdigitalhistory.org/en/article/jXupS3QAeNgb

Leurs, K. (2017). Feminist data studies: Using digital methods for ethical, reflexive and situated socio-cultural research. *Feminist Review*, *115*(1), 130-154.

Moretti, Franco. (2013) Distant Reading. London, New York: Verso.

Pašeta, S. (2017). Feminist Political Thought and Activism in Revolutionary Ireland, c. 1880–1918. *Transactions of the Royal Historical Society*, *27*, 193-209.

Walsh, F. (2020). *Irish Women and the Great War*. Cambridge University Press.

Notes

1. (1) Examine power; 2) Challenge power; 3) Elevate emotion and embodiment; 4) Rethink binaries and hierarchies; 5) Embrace pluralism; 6) Consider context; 7) Make labor visible)

The Agency of Brokerage: Reading Positioning and Power in Character Networks

Selisker, Scott
selisker@arizona.edu
University of Arizona, United States of America

This paper considers brokerage as a key interpretive aspect of character networks within narratives. Automatically extracted co-occurrence networks find brokerage through relatively high betweenness centrality. Brokerage anchors interpretations of various forms of social power, here explored through the grassroots political networks of Karen Tei Yamashita's 2010 novel *I Hotel* (Yamashita, 2010).

In this work, I have been developing and building on models for automatic extraction of social networks, while at the same time developing interpretive methods for considering a major feature of networks: brokerage and betweenness centrality (Burt, 2005). I use Bamman et al's BookNLP package to extract and correct characters and organizations and pronominal mentions of them to exract a highly accurate co-occurrence network, then discuss findings on betweenness centrality relative to characters' roles within the text.

For my project, network analysis allows us to interpret and think closely about how figures' positioning between groups becomes meaningful within fiction: as a point for identifying go-between figures like native informants and racial passers, as a model of privacy within fictional narrative, and for considering the relationships between community and identity (Selisker, 2018). Within fictional narratives, conventional measures of centrality used in other fields' implementations of social network analysis are seldom immediately meaningful; rather, following and building on work by Algee-Hewitt and Sims and Bamman, betweenness centrality and the flow of information allow us to see how positioning matters within character networks (Sims and Bamman, 2020; Algee-Hewitt, 2017). Access to information, and privileged access to other groups, is both a major thematic feature of many narratives that consider segregated social spaces, access to power, and more, and an aspect of a text that is readily visible in the highly abstract "slice" of information that a co-occurrence network can provide.

Yamashita's *I Hotel*, a novel about a multiracial coalition of Asian Americans in 1970s San Francisco, is both thematically and structurally a demonstration of the agency of brokerage: the heroes of the text are those who build out the coalition, both among Asian American progressive communities and with other groups. The non-hierarchical collectives described in the novel, based on meticulous archival research and interviews by Yamashita, resemble the non-hierarchical, and decentralized network forms explicitly espoused by many political movements betwen the 1970s and the present (the Combahee River Collective, Zapatistas, Occupy Wall Street, Black Lives Matter).

We see, by using network extraction and the methods of social network analysis, that the text's explicitly non-hierarchical collectivity is anything but formless, and that the social network is by no means the freeform opposite to the restraints of institutions. Social network analysis as a reading practice allows us to discover and consider brokerage as a textual theme: the network "protocols" that hold the coalition together, the labor involved in coalition-building, and the importance of broker figures to the novel's project as a whole. High betweenness centrality relative to other metrics (degreee, eigenvector centrality) offers a tool for highlighting go-between roles within the text, and for seeing the structure of the coalition at the center of the novel. These are the concrete features of networked collectivity as we imagine its ideals, and the question of brokerage as a form of agency can allow us to explore and compare how this ideal functions in this and other contemporary fiction that addresses diversity and identity.

Bibliography

Algee-Hewitt, M. (2017). Distributed Character: Quantitative Models of the English Stage, 1550–1900. *New Literary History*, **48**(4): 751–82.

Burt, R. S. (2005). *Brokerage and Closure: An Introduction to Social Capital*. New York: Oxford University Press.

Jagoda, P. (2016). *Network Aesthetics*. Chicago: University of Chicago Press.

Le-Khac, L. (2020). *Giving Form to an Asian and Latinx America*. Stanford: Stanford University Press.

Moretti, F. (2011). Network Theory, Plot Analysis. *New Left Review*, **68**: 80–102.

Schantz, N. (2008). *Gossip, Letters, Phones: The Scandal of Female Networks in Film and Literature*. New York: Oxford University Press.

Selisker, S. (2018). The Novel and WikiLeaks: Transparency and the Social Life of Privacy. *American Literary History*, **30**(4): 756–76.

Sims, M. and Bamman, D. (2020). Measuring Information Propagation in Literary Social Networks. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 642–52.

Smeets, R. (2021). *Character Constellations: Representations of Social Groups in Present-Day Dutch Literary Fiction*. Leuven: Leuven University Press.

So, R. J. and Long, H. (2013). Network Analysis and the Sociology of Modernism. *Boundary 2*, **2**(40): 2.

Stier, A. (2013). Ties That Bind: American Fiction and the Origins of Social Network Analysis. *ProQuest Dissertations Publishing*.

Woloch, A. (2003). *The One Vs. The Many: Minor Characters and the Space of the Protagonist in the Novel*. Princeton: Princeton University Press.

Wong, L. (2017). Dwelling over China: Minor Transnationalisms in Karen Tei Yamashita's I Hotel. *American Quarterly*, **69**(3): 719–39.

Yamashita, K. T. (2010). *I Hotel*. Minneapolis, MN: Coffee House Press.

Genre Classification in English Poetry with Lexical and Prosodic Features

Shang, Wenyi

wenyis3@illinois.edu
School of Information Sciences, University of Illinois at Urbana-Champaign, United States of America

Underwood, Ted

tunder@illinois.edu
School of Information Sciences, University of Illinois at Urbana-Champaign, United States of America; Department of English, University of Illinois at Urbana-Champaign, United States of America

There has been an abundance of research on the lexical (Bergel et al., 2016) and prosodic (Anttila and Heuser, 2016) features of poetic texts. Some recent attempts such as Šeļa et al. (2020) combine the two feature sets to model the association between poetic meter and meaning. In this project, we ask how the relative importance of lexical and prosodic features vary across different forms and genres. We pose this question by classifying different categories of English poetry with lexical and prosodic features separately. Then we compare the results of the parallel experiments, to ask which categories have a stronger lexical or prosodic character.

We collected all 37,700 English poems that are tagged with a certain "genre" from the Chadwyck-Healey Literature Collections (http://collections.chadwyck.com). (Note that our use of the term "genre" is drawn from our source; scholars might well characterize some of these genres as "forms" or "modes.") Since many of the 44 total genres only have a few cases, we kept the poems of the top 8 genres only, which consist of 30,704 (81.44%) poems. Next, we trained a lexical classifier and a prosodic classifier and extracted features for each of them. For the lexical classifier, the features were extracted by transforming the texts into an array of word frequencies, and a grid search was conducted to select the algorithm and number of features used for classification. Optimization is achieved when the random forest algorithm and top 500 features are used. As for the prosodic classifier, we used the Python library "Poesy" (https://github.com/quadrismegistus/poesy) to obtain "foot type", "feet number", and "rhyme style" of each poem. Once again, the best performance is achieved when the random forest algorithm is used.

We then conducted classification experiments based on a random train-test split (70% training vs 30% testing) and compared the performances of the two classifiers on an 8-genre classification task and a 2-class classification task, where we classified each of the top 8 forms against all others (e.g., "sonnet" vs "non-sonnet", "ballad" vs "non-ballad").

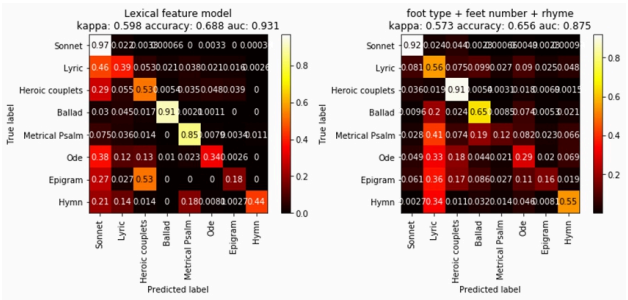


Figure 1.
Confusion matrix of models trained with different feature sets (8-genre classification)

In the 8-genre classification experiment, while the lexical and the prosodic classifier have similar overall

performances, their results on different genres significantly differ: the lexical classifier classifies ballads and metrical psalms very well, but easily confuses heroic couplets with sonnets and epigrams, while the prosodic classifier classifies heroic couplets very well but easily confuses ballads and metrical psalms with lyrics.

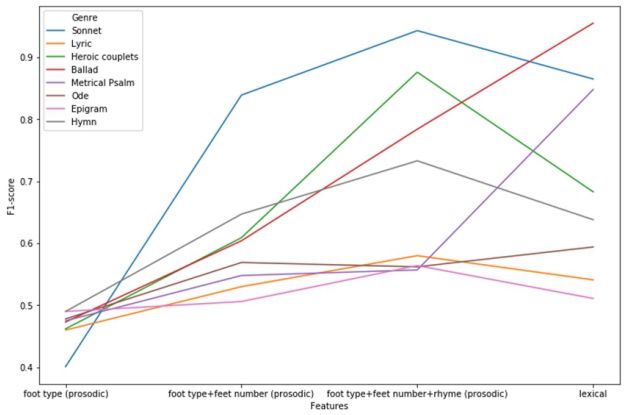


Figure 2.
F1 score of the classification experiments of different genres trained with different feature sets (2-class classification)

The results of 2-class classification demonstrate a similar pattern: sonnets and heroic couplets are better distinguished from other forms when using classifiers trained with prosodic features, while ballads and metrical psalms are better distinguished from other forms when using classifiers trained with lexical features. These results preliminarily suggest that different genres are distinguished with different features: while ballads and metrical psalms are distinguished by the diction they use, sonnets and heroic couplets are defined by prosodic features.

We also examined the performance of both classifiers on poems by different authors. Here, we used the "hold-out-one" strategy: we selected all poems by authors with at least 30 poems in the dataset and trained the classifiers with all poems except those written by an author, and tested the results on the poems written by that author.

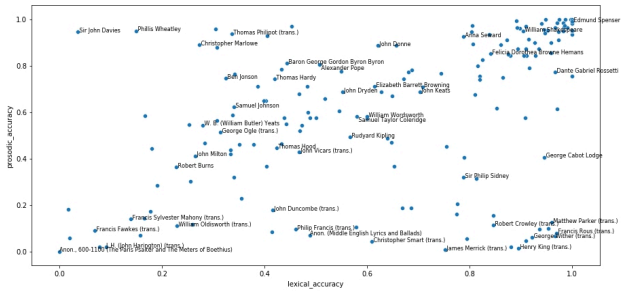


Figure 3.
Performances of classifiers on poems of different poets

In figure 3, most authors of translated works are in the bottom of the figure (their works cannot be easily classified by prosodic features). We also see preliminary evidence that there might be a correlation between poetic prominence and prosodic regularity: the prominent poets mostly appear in the upper half (their works can be well classified by prosodic features). It is likely that the prominent poets continuously stick to certain prosody conventions for each genre, while the translated poems use very different prosody in the same genre, which is understandable as prosodic pattens are easily lost in translation. The accuracy of classification with lexical features does not show a consistent pattern. However, two influential early modern authors, Spencer and Shakespeare, appear in the top-right corner (their works can be very well classified by both lexical and prosodic features), perhaps indicating that they set the standard for following authors in both the diction and prosody used in different genres.

To further explore these observations, in the future we will further investigate the relationships between the accuracy of prediction and factors such as the date of the poem, the poets' canonicity, and their nationality. Additionally, we will also take word order into consideration and use N-gram in supplement of single-word frequencies.

The findings above already tell us a lot of things about the history of English poetry like the roles of lexical and prosodic features in poetry genre classification, and how such roles differ in different genres. However, the most important potential contribution of this project is to distinguish the concepts of "genre" and "form": if some poetic categories are best identified by prosody, and others by "content" (represented by lexical features), it might be possible to disentangle "form" from other aspects of "genre." For example, it is observed that heroic couplets and epigrams use similar diction, what distinguish them as two "genres" is their "forms"; in contrast, ballads and lyrics are in similar "forms", and they are considered as different "genres" because of the words they use.

This could reinforce the argument of King (2021) that "genre" operates on a larger scale than "form". Furthermore, in addition to understanding how the transformation of poetic "forms" was related to narratives of culture (Martin, 2012), insights can also be gained on the evolution of the roles of "content" and "form" in defining "genres" of poems by different poets and in different periods and locations. While various follow-up experiments need to be done, there is preliminary evidence that the works of prominent poets tend to be close to a prosodic prototype for a genre.

Bibliography

Anttila, A. and Heuser, R. (2016). Phonological and Metrical Variation across Genres. In Hansson, G. Ó., Farris-Trimble, A., McMullin, K. and Pulleyblank D. (eds), *Proceedings of the 2015 Annual Meetings on Phonology*. Washington, D.C.: Linguistic Society of America. https://doi.org/10.3765/amp.v3i0.3679

Bergel, G., Howe, C. J. and Windram, H. F. (2016). Lines of Succession in an English Ballad Tradition: The Publishing History and Textual Descent of The Wandering Jew's Chronicle. *Digital Scholarship in the Humanities*, **31**(3), 540–562.

King, R. S. (2021). The Scale of Genre. *New Literary History*, **52**(2), 261–284. https://doi.org/10.1353/nlh.2021.0012

Martin, M. (2012). *The Rise and Fall of Meter: Poetry and English National Culture, 1860– 1930*. Princeton: Princeton University Press.

Šeļa, A., Orekhov, B. and Leibov, R. (2020). Weak Genres: Modeling Association Between Poetic Meter and Meaning in Russian Poetry. In Karsdorp, F., McGillivray, B., Nerghes, A. and Wevers, M. (eds), *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)*. CEUR Workshop Proceedings, pp. 12–31.

Structural Balance in the Historical Political Networks of China

Shang, Wenyi
wenyis3@illinois.edu
School of Information Sciences, University of Illinois at Urbana-Champaign, United States of America

Chen, Song
song.chen@bucknell.edu
Department of East Asian Studies, Bucknell University, United States of America

Negative ties, especially when studied in tandem with positive ones, are instrumental for understanding political culture in history. How likely was it for political adversaries to have an amicable relationship with some third party? Were enemies of an enemy always friends? Different answers to these questions reflect different political cultures and dynamics of each given moment in history. Although recent works on premodern Chinese political networks have benefited immensely from the method of social network analysis, they have focused mainly on positive relations and rarely looks at negative political ties (e.g., Chen, 2016). The only known study in the field that takes negative ties into consideration (Yan & Wang, 2018) investigates merely the proportion of balanced and unbalanced triadic structures, but fails to provide a more nuanced discussion of the historical significance of different types of balanced or unbalanced triads.

This study studies triads in historical Chinese political networks by applying the concept of structural balance (Heider, 1946; Cartwright & Harary, 1956) to three moments of Chinese history. Following the structural balance theory, we classify the triads in an undirected signed graph into four distinct types based on the number of positive ties in each triad, and a triad with an odd number of positive ties is considered "balanced":

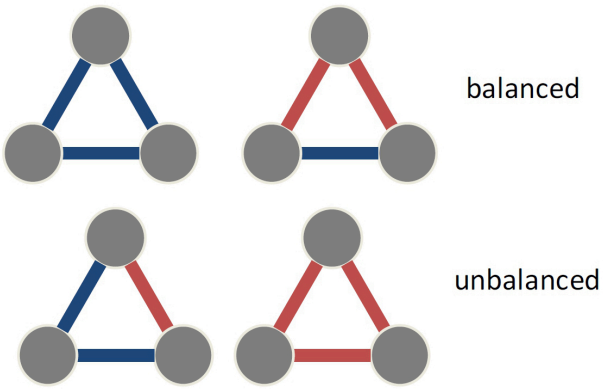


Figure 1.
Structural balance (positive ties are in blue, and negative ties in red)

We further interpret each type of triad as follows:

1. A triad where all ties are positive, denoted by "+++", indicates "political collegiality" (two actors who have a shared friend are also friends with each other);
2. A triad with two positive ties and one negative tie, denoted by "++–", indicates "political tolerance" (two enemies nevertheless are both friends with the same third party);
3. A triad with one positive tie and two negative ties, denoted by "+––", indicates "political polarization" (two actors who have a shared enemy are friends with each other);
4. A triad where all ties are negative, denoted by "–––", indicates "political plurality" (two actors who have a shared enemy nevertheless fight between themselves).

We applied the above concepts to the sociopolitical networks in three periods of Chinese history: Eastern Jin (317–420), the Qingli (1041–1048) and Xining-Yuanfeng

eras (1068–1085, hereinafter XiFeng) of the Song dynasty (960–1279). Eastern Jin marks the peak of aristocratic dominance when political actors came preponderantly from a small number of status-conscious, endogamous "great clans" (Lewis, 2009, p. 51), whereas the two periods of Song are both known for political reforms and attendant factional struggles. Network data on the Eastern Jin political elite were extracted from A New Account of the Tales of the World (Liu, 2002), a collection of anecdotes providing the richest extant account of elite interactions in the fourth century. Here, whenever a single anecdote demonstrates a relationship between two people, a tie is created between them and is manually classified as positive (e.g., bureaucratic appointment, gift-giving) or negative (e.g., attacking, hostility). Data of positive and negative ties on the two eras of the Song are exported from China Biographical Database (CBDB) (Harvard University et al., 2021), supplemented by negative ties manually gleaned from the section on bureaucratic dismissals and demotions in Song huiyao jigao [Collected Administrative Documents of the Song] (Xu, 1957). In all three networks, it is not rare for a pair of actors to have relationships of different nature at different times and are thus connected by both positive and negative ties. Each of these dyads is counted twice, first as one linked by a positive tie and then by a negative tie.

Our results show that the Song-Qingli network has the highest triadic balance score (0.79), followed by the Song-XiFeng network (0.63), whereas the Eastern Jin network has the lowest (0.53). The proportion of dyads with a mix of positive and negative ties in each network has a reversed rank: highest in Eastern Jin (8.7%), followed by the Song-XiFeng (2.3%) and the Song-Qingli (0.8%) networks. Both calculations suggest—in congruence with findings in the existing scholarship (e.g., Zhu 1985)—that the political world in Eastern Jin was the most fluid and that political actors in Eastern Jin could not be classified into two neatly demarcated opposing groups.

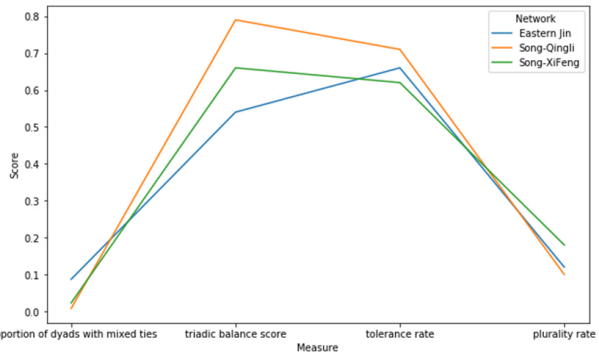


Figure 2.
Measurements of the three different networks

A comparison between two Song networks also paints a more complicated picture of eleventh-century court politics than what the conventional narrative of reformers versus conservatives has made us believe (e.g., Liu, 1959). On the one hand, a higher degree of political tolerance in the Qingli era, measured by "++−" triads as a percentage of all triads with one or more negative ties (0.71 in Song-Qingli vs. 0.62 in Song-XiFeng), indicates a higher likelihood for political adversaries to have common friends. On the other hand, the Song-XiFeng network has a higher plurality score (computed as the percentage of the "−−−" triads in all triads with two or more negative ties) than the Song-Qingli network (0.18 in Song-Xifeng vs. 0.10 in Song-Qingli), implying a more pluralistic and dynamic political scene in the Xining-Yuanfeng era characterized by shifting alliances and more frequent discord among both reformers and their opponents.

These explorations generate a fresh insight into pre-modern Chinese political culture to be pursued in greater depths in future work. We plan to improve the quality of our data, employ new analytical methods (e.g., balance theory-based blockmodeling), and combine it with more conventional textual studies.

Bibliography

Cartwright, D., & Harary, F. (1956). Structural Balance: A Generalization of Heider's Theory. *Psychological Review*, **63**(5), 277–293.

Chen, S. (2016). Governing a Multicentered Empire: Prefects and Their Networks in the 1040s and 1210s. In Ebrey, P. B. and Smith, P. J. (eds), *State Power in China: 900–1325*. Seattle: University of Washington Press, pp. 101–152.

Harvard University, Academia Sinica, and Peking University (2021). *China Biographical Database (CBDB)*. https://projects.iq.harvard.edu/cbdb (accessed 10 December 2021).

Heider, F. (1946). Attitudes and Cognitive Organization. *The Journal of Psychology*, **21**(1), 107–112.

Lewis, M. E. (2009). *China Between Empires: The Northern and Southern Dynasties (Vol. 2)*. Cambridge and London: The Belknap Press of Harvard University Press.

Liu, I. (2002). *Shih-shuo Hsin-yü: A New Account of Tales of the World* (R. B. Mather, Trans.; Second Edition). Ann Arbor: Center for Chinese Studies, The University of Michigan.

Liu, J. (1959). *Reform in Sung China: Wang An-shih (1021–1086) and His New Policies*. Cambridge: Harvard University Press.

Xu, S. (1957). *Song huiyao jigao* [Collected Administrative Documents of the Song]. Beijing: Zhonghua shuju.

Yan, C., & Wang, J. (2018). Shuzi renwen shijiao: jiyu fuhao fenxi fa de songdai zhengzhi wangluo keshihua yanjiu [Digital Humanistic Perspective: A Study on the Visualization of Political Network in Song Dynasty Based on Symbolic Analysis]. *Zhongguo Tushuguan Xuebao*, (5), 87–103.

Zhu, Z. (1985). Shilun Dongjin houqi gaoji shizu zhi moluo ji Huan Xuan dai Jin zhi xingzhi [On the Decline of the High Aristocratic Families and the Nature of Huan Xuan's Replacement of Jin in the Late Eastern Jin]. *Beijing daxue xuebao (zhexue shehuikexue ban)*, (3), 77–90.

Developing a text readability system for Sesotho based on classical readability metrics

Sibeko, Johannes

johannes.sibeko@mandela.ac.za
Nelson Mandela University, South Africa

Van Zaanen, Menno

menno.vanzaanen@nwu.ac.za
South African Centre for Digital Language Resources, South Africa

Sesotho as one of South Africa's (SA) eleven official languages is a home language to about eight percent of SA inhabitants and 98 percent of the population in Lesotho (Reid et al., 2019). Like many Asian languages, Sesotho is an under-resourced language (Wills et al., 2020). The repository of the South African Centre for Digital Language Resources (SADiLaR) provides the limited Sesotho resources (see https://repo.sadilar.org/).

This project aims to develop a readability tool for Sesotho texts. When additional language resources are required, these will also be developed. For readers (especially learners) to select texts suitable for their reading level, a measure of readability for texts is essential.

Existing text readability investigations in the context of SA, have focused mainly on health documents (Joubert and Githinji, 2014; Krige and Reid, 2017; Leopeng, 2019; De Wet, 2021) and textbooks (Sibanda, 2013; Wissing et al., 2016). Krige and Reid (2017) used three English metrics to measure readability of medical pamphlets in Sesotho, which does not consider differences between the languages. Language specific readability metrics should be developed before proper conclusions can be drawn. To our knowledge, no language specific readability metrics have exist for any African language, apart from Afrikaans (Jansen et al., 2017). Unfortunately, no implementations of these metrics could be found.

To develop readability metrics, texts with known readability levels are needed. Unfortunately, for Sesotho, copyright restrictions limit access to texts with (expected) known readability levels, such as textbooks. However, in SA, Sesotho is tested at high school on two levels, home language (HL) and first additional language (FAL). We expect these exam texts to have consistent readability over the years, with HL texts more difficult to read than FAL texts. To test this, we analyzed the readability of SA English HL and FAL exam texts (Sibeko and Zaanen, 2021) using existing metrics, which showed that the readability of the texts is consistent over time and different between the two levels.

If we assume that the development of the exam texts for Sesotho (and the other SA languages) follows the same process as that for English texts (Sibeko and Zaanen, 2021), Sesotho exam texts also show clear differences in levels of readability. They can then be used for the development of readability metrics for Sesotho.

We currently build on text properties used in nine readability metrics for English (Sibeko and Zaanen, 2021): Flesch-Kincaid Grade Level (Kincaid) (Kincaid et al., 1975), Flesch Reading Ease (Flesch) (Flesch, 1948), Simple Measure of Gobbledygook (SMOG) (Mc Laughlin, 1969), Gunning Fog index (Fog) (Gunning, 1952; Gunning, 1969), läsbarhetsindex (LIX) (Björnsson, 1968), Rate index (RIX) (Anderson, 1983), Automated Readability index (ARI) (Senter and Smith, 1967; Kincaid and Delionbach, 1973), Coleman-Liau index (Coleman and Liau, 1975), and the Dale-Chall index (Dale and Chall, 1948).

The readability metrics rely on text properties such as word and sentence length. Due to differences in language structure, these properties cannot be applied readily to other languages. To this end, we are re-conceptualising properties, such as long words, which have more than six characters in the LIX and RIX metrics, difficult words, which do not appear in the 3000 most frequently used English words in the Dale-Chall Index, and complex words, which have more than two syllables in the Gunning Fog Index, to reflect Sesotho's context. In particular, features such as syllables and frequently used words are language specific.

To resolve these issues, we currently develop automated Sesotho syllabification systems, including a rule-based system based on Guma's (1982) description and a pattern-based system (using TeX's hyphenation system (Liang, 1983)). Additionally, we investigate the concepts of long, difficult, and complex words in Sesotho. To make matters more complex, Sesotho has two orthographies, Lesotho (LS) and SA orthography (SAS) (Motjope-Mokhali et al., 2020).

We currently use SAS orthography given the usage of the SA high school exam texts.

Once the different text properties are defined, they can be applied to the Sesotho exam texts. The values can then be combined in linear regression models, which will provide mathematical formulas that provide a level of text readability for Sesotho texts.

This contribution describes progress in the development of the first automated text readability analysis tool for a SA language (Sesotho). Given the limited availability of computational resources for Sesotho, we also describe language resources developed within the project. To aid the development of digital language resources, all developed Sesotho resources will be published in open repositories, such as Github and SADiLaR's repository.

Bibliography

Anderson, J. (1983). Lix and rix: Variations on a little-known readability index. *Journal of Reading*, **26**(6): 490–96.

Björnsson, C. (1968). *Läsbarhet.* (Pedagogiskt Utvecklingsarbete Vid Stockholms Skolor. 6). Liber: Solna, Seelig.

Coleman, M. and Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, **60**(2): 283–84.

Dale, E. and Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational Research Bulletin*: 37–54.

De Wet, A. (2021). The development of a contextually appropriate measure of individual recovery for mental health service users in a South African context Stellenbosch University, Stellenbosch, South Africa PhD thesis.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, **32**(3): 221.

Guma, S. (1982). *An Outline Structure of Southern Sotho*. 2nd ed. Pietermaritzburg, South Africa: Shooter; Shuter Publishers.

Gunning, R. (1952). *Technique of Clear Writing*. McGraw-Hill.

Gunning, R. (1969). The fog index after twenty years. *Journal of Business Communication*, **6**(2): 3–13.

Jansen, C., Richards, R. and Van Zyl, L. (2017). Evaluating four readability formulas for Afrikaans. *Stellenbosch Papers in Linguistics Plus*, **53**: 149–66.

Joubert, K. and Githinji, E. (2014). Quality and readability of information pamphlets on hearing and paediatric hearing loss in the gauteng province, South Africa. *International Journal of Pediatric Otorhinolaryngology*, **78**: 354–58.

Kincaid, J. P. and Delionbach, L. J. (1973). Validation of the automated readability index: A follow-up. Human Factors, **15**(1): 17–20.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L. and Chissom, B. S. (1975). *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Naval Technical Training Command Millington TN Research Branch.

Krige, D. and Reid, M. (2017). A pilot investigation into the readability of sesotho health information pamphlets. *Communitas*, **22**: 113–23.

Leopeng, M. T. (2019). Translations of informed consent documents for clinical trials in South Africa: Are they readable? University of Cape Town, Cape Town, South Africa Master's thesis.

Liang, F. (1983). Word hy-phen-a-tion by com-put-er Stanford, USA: Stanford University PhD thesis.

Mc Laughlin, G. H. (1969). SMOG grading-a new readability formula. *Journal of Reading*, **12**(8): 639–46.

Motjope-Mokhali, T., Kosch, I. and Mafela, M. J. (2020). Sethantso sa sesotho and Sesuto-English dictionary: A comparative analysis of their designs and entries. *Lexikos*, **30**: 1–17.

Reid, M., Nel, M. and Janse van Rensburg-Bonthuyzen, E. (2019). Development of a Sesotho health literacy test in a South African context. *African Journal of Primary Health Care and Family Medicine*, **11**(1): 1–13.

Senter, R. and Smith, E. A. (1967). *Automated Readability Index*. Cincinnati University, OH.

Sibanda, L. (2013). A case study of the readability of two grade 4 natural sciences textbooks currently used in South African schools Rhodes University, Grahamstown, South Africa Master's thesis.

Sibeko, J. and Zaanen, M. van (2021). An analysis of readability metrics on English exam texts. In, *Proceedings of the International Conference of the Digital Humanities Association of Southern Africa (Dhasa)*.

Wills, S., Uys, P., Heerden, C. J. van and Barnard, E. (2020). Language modeling for speech analytics in under-resourced languages. In, *Proceedings of the 21st Annual Conference of the International Speech Communication Association (Interspeech 2020), Shanghai, China*. International Speech Communication Association, pp. 4941–45.

Wissing, G.-J., Blignaut, A. S. and Van den Berg, K. (2016). Using readability, comprehensibility and lexical coverage to evaluate the suitability of an introductory accountancy textbook to its readership. *Stellenbosch Papers in Linguistics*, **46**: 155–79.

Rule-based Speaker Identification for Speech, Thought and Writing in German Literary Texts

Sluyter-Gäthje, Henny
sluytergaeth@uni-potsdam.de
University of Potsdam

1. Introduction

To study storyworlds created in literary texts, the analysis of characters and their interaction is one of the most fundamental aspects. A character's voice in a storyworld can be expressed by speech, thought or writing (STW), the representation of which can take on different forms depending on how truthful it is to the original utterance (Genette, 1983: 171-173). The following types can be differentiated: direct (most truthful), indirect, reported (least truthful) and free indirect (mixture of direct and indirect) STW.

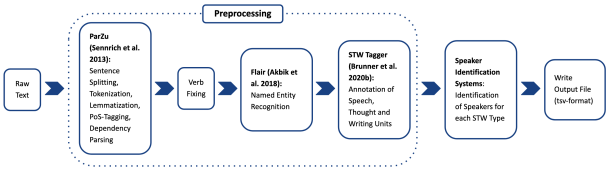
A basis for the automatic processing of storyworlds is the identification of STW units and the attribution to their producers, i.e. the characters. While successful approaches for the recognition of STW units do exist, most speaker attribution systems are limited to direct or indirect speech. In this work, we develop rule-based speaker identification systems for the attribution of not only speech but also thought and writing, not limited to direct and indirect but also including reported and free indirect representations in German literary texts.

2. Related Work

The task of speaker attribution can be divided into two subtasks: the identification of speakers –finding the textual mention of a speaker– and the resolution of speakers – resolving the textual mention to a speaker entity. This work is concerned with speaker identification. Early approaches to speaker attribution mostly relied on pattern matching (e.g. Krestel et al., 2008). Elson and McKeown (2010) presented a first machine learning (ML) approach which formed the basis for follow-up work (direct: O'Keefe et al., 2012; He et al., 2013; Yeung and Lee, 2017; Ek et al., 2018; indirect: Pareti et al., 2013; Newell et al., 2018). Krug et al. (2016) experimented with a rule-based approach for the attribution of direct speech units in German literary texts which could outperform their ML approaches. Similarly, Muzny et al. (2017) built a state-of-the-art system for the domain of English literature which attributed speakers for direct speech in a rule-based way.

3. Approach

This work builds on the approaches of Krug et al. (2016) and Muzny et al. (2017) by adapting and extending the rules they presented. Additionally, we formulate new rules. All rules are compiled, manually evaluated and improved in an iterative way with the help of the Corpus Redewiedergabe (Brunner et al., 2020a). For each representation type (direct, indirect, reported and free indirect) we build one system with a different set and a different order of rules; a rule can only be applied once. Similar to related work, our systems rely heavily on linguistic annotations (see figure 01) and on predefined word lists (e.g. to identify animate nouns). A final evaluation was performed on a held-out test set extracted from the Corpus Redewiedergabe. The full pipeline of the systems, including the recognition of STW units (Brunner et al., 2020b), is shown in figure 01. The systems are publicly available alongside an extensive description of the rules.



Pipeline of the speaker identification systems for the annotation of raw text.

4. Results

Author	STW type	Performance Range	STW medium	Domain	Language
Pareti et al. (2013)	Direct Indirect Mixed	**85 – 91** 74 – 79 65 – 81	Speech	News	English
Krug et al. (2016)	Direct	78.4	Speech	Literature	German
Muzny et al. (2017)	Direct	76 – 85	Speech	Literature	English
This work with gold STW annotations	Direct Indirect Reported Free indirect	63.91 82.2 71.38 50.0	Speech, Thought, Writing	Literature	German

Comparison of the accuracies of speaker attribution systems that were used in setups comparable to this work. Accuracy ranges are indicated as some systems were applied to different data sets with varying success. Maximum values are marked in bold.

As shown in figure 02 our systems achieve the best performance for attributing indirect, reported and free indirect STW. The direct system could be improved, for example when handling conversational patterns. The full pipeline achieves a comparable performance.

5. Future Work

The pipeline itself could be improved (e.g. by extending the predefined word lists) and the systems could be tested on and eventually adapted to another domain. For comparative purposes, neural networks that use semantic word representations could be trained for the task of speaker identification. Finally, the systems could be extended to also resolve speakers. The systems can be used as is to perform analyses in the field of Computational Literary Studies e.g. to address gender related research questions (cf. Schumacher and Flüh, 2020).

Bibliography

Akbik, A., Vollgraf, R. and Blythe, D. (2018). Contextual String Embeddings for Sequence Labeling. In *27th International Conference on Computational Linguistics*. COLING 2018. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 1638–49.

Brunner, A., Engelberg, S., et al. (2020). Corpus REDEWIEDERGABE. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC'20)*. LREC. Marseille, France: European Language Resources Association, pp. 803–12.

Brunner, A., Duyen, N., et al. (2020). To Bert or Not to Bert–Comparing Contextual Embeddings in a Deep Learning Architecture for the Automatic Recognition of Four Types of Speech, Thought and Writing Representation. In *Proceedings of the 16th Conference on Natural Language Processing (KONVENS 2020)*. Konvens. Zurich, Switzerland.

Ek, A. et al. (2018). Identifying Speakers and Addressees in Dialogues Extracted from Literary Fiction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. LREC. Miyazaki, Japan.

Elson, D. and McKeown, K. (2010). Automatic Attribution of Quoted Speech in Literary Narrative. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*. AAAI. AAAI Press, pp. 1013–9.

Genette, G. (1990). *Narrative Discourse: An Essay in Method*. 1. publ., 4. print. Ithaca: Cornell University Press.

He, H., Barbosa, D. and Kondrak, G. (2013). Identification of Speakers in Novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Buldaria: Association for Computational Linguistics, pp. 1312–20.

Krestel, R., Bergler, S. and Witte, R. (2008). Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association.

Krug, M. et al. (2016). Attribuierung Direkter Reden in Deutschen Romanen Des 18.-20. Jahrhunderts. Methoden Zur Bestimmung Des Sprechers Und Des Angesprochenen. In *DHd 2016, Modellierung - Vernetzung - Visualisierung, Die Digital Humanities Als Fächerübergreifendes Forschungsparadigma, Konferenzabstracts*. 124-130. DHd. Leipzip, Germany.

Muzny, F. et al. (2017). A Two-Stage Sieve Approach for Quote Attribution. In *In Proceedings of the 15th Conference of the European Chapter of the Association for Computation al Linguistics*. Valencia, Spain: Association for Computational Linguistics, pp. 460–70.

Newell, C., Cowlishaw, T. and Man, D. (2018). Quote Extraction and Analysis for News. In *Proceedings of KDD Workshop on Data Science Journalism and Media (DSJM)*. New York, NY, USA: Association for Computing Machinery.

O'Keefe, T. et al. (2012). A Sequence Labelling Approach to Quote Attribution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea: Association for Computational Linguistics, pp. 790–9.

Pareti, S. et al. (2013). Automatically Detecting and Attributing Indirect Quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, pp. 989–99.

Schumacher, M. and Flüh, M. (2020). M*w Figurengender Zwischen Stereotypisierung Und Literarischen Und Theoretischen Spielräumen: Genderstereotypen Und -Bewertungen in Der Literatur Des 19. Jahrhunderts. In *DHd 2020, Spielräume, Digital Humanities Zwischen Modellierung Und Interpretation, Konferenzabstracts*. Paderborn, Germany, pp. 162–6.

Sennrich, R., Volk, M. and Schneider, G. (2013). Exploiting Synergies between Open Resources for German Dependency Parsing, Pos-Tagging, and Morphological Analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. Shoumen, Bulgaria: INCOMA Ltd, pp. 601–9.

Yeung, C. Y. and Lee, J. (2017). Identifying Speakers and Listeners of Quoted Speech in Literary Works. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*. Taipei, Taiwan: Asian Federation of Natural Language Processing, pp. 325–9.

Semantic Data Lakes for Knowledge Extraction in the Humanities: A Case Study on Bernard Berenson's Network of Acquaintances

Spinaci, Gianmarco
gspinaci@itatti.harvard.edu
I Tatti - The Harvard University Center for Italian Renaissance Studies, Italy

Grillo, Remo
grilloremo@gmail.com
I Tatti - The Harvard University Center for Italian Renaissance Studies, Italy

Klic, Lukas
lklic@itatti.harvard.edu
I Tatti - The Harvard University Center for Italian Renaissance Studies, Italy

Bonora, Paolo
paolo.bonora@unibo.it
Alma Mater Studiorum - Università di Bologna

Problem description and research question:

It is frequent practice in Digital Humanities (DH) studies to create Knowledge Bases (KB) limited to specific domains of interest. This leads to the creation of a plethora of highly specialized data silos [1]. As a result, the extraction of knowledge dispersed across KBs can be challenging. We argue that an extended KB — a Semantic Data Lake [2] (SDL) — obtained by aggregating heterogeneous vertical KBs could help define a set of heuristics to support transversal Knowledge Extraction [3]. We propose a case study based on a set of vertical KBs with converging content based on the correspondence of Bernard Berenson (1909-1960), the diaries of Mary Berenson (1879-1935), and their personal photographic archive.

As these semantic silos contain converging content, we aim to demonstrate that their union could support the discovery of original information. Through inductive reasoning, we aim to analyze data looking for graphs in order to assess the likelihood of a relationship among two or more actors. We aim to reconstruct a network of acquaintances by analyzing these paths within a consolidated dataset. If the proposal will be effective in identifying and qualifying the relationships among actors in the Berenson Circle, Art historians and domain experts will evaluate the relevance of these relationships.

Methodology:

The data lake-based approach does not require a prior alignment of different ontologies within KBs. In other words, sources do not necessarily need to use the same ontological framework or re-use the same modelling patterns, as long as each source is modeled consistently. KBs are then selected exclusively on the basis of their contents. The only prerequisite is that they are represented in RDF. In order to take into account a broader set of available KBs, we adopted an ontology-agnostic methodology. Knowledge Extraction begins with the analysis of the structure of the paths connecting target entities, in our case Berenson's acquaintances. This requires a reconciliation of actor identifiers and harmonized space/time dimensions. Regarding the entities that have not been already qualified, we used a NER algorithm to extract them and, together with qualified entities, we used Wikidata for a semi-supervised reconciliation process.

Paths are then identifiable through the data lake's composite graph, both from a quantitative (i.e. shortest paths) and a qualitative (i.e. semantics of the resulting paths' graph patterns) approach. Among this set of identified paths, the most relevant ones will be selected for answering the research question.

The extraction process will be organized into the following steps:

1. The SDL is built from source KBs partitioned within their own named graph
2. Reconciliation of target entities (i.e. people);
3. Harmonization of space/time coordinates (i.e. toponyms disambiguation and georeferencing);
4. Extraction of paths between instances of people;
5. Analysis and selection of paths that allow for relationships among people to be inferred;
6. Formalization of these paths as SPARQL Property Paths or SWRL rules;
7. Validation of extracted paths across KBs;

Sources:

This case study is designed around a strong convergence of sources with a known set of interactions between actors across space and time through common references to events, people, and places. Sources comprise letters exchanged between Berenson and Yukio Yashiro [4] (115 texts written from 1925 to 1960), letters sent from Belle Da Costa Greene to Berenson (470 texts written from 1909 to 1949), diaries of Mary Berenson (30 annual diaries written from 1879 to 1935), and metadata from historical photographs housed in the Berenson archive. We are focusing on the 1925 to 1935 period when we have the best overlapping of the corpora.

Expected results and validation:

In the correspondence between Yukio Yashiro and Bernard Berenson, numerous references are made regarding meetings with unspecified art historians. Diary entries by Mary Berenson report the guests at their residence for most days of the year, including Yashiro. By cross-referencing these sources, we can reconstruct a network of acquaintances between Yashiro, the Bereson's, and others. The numbers of extracted entities are: more than 1k names tagged as persons of which 20 qualified, and almost 300 toponyms of which 150 qualified. Once this information has been extracted through the proposed methodology and formalized in new assertions, new validation criteria should be adopted in order to refine the data quality and assess their effectiveness. These criteria need to consider metrics such as frequency count, recall and precision, and accuracy analysis by reconciliation with sources.

Conclusions:

We presented a SDL methodology that allows for a lightly supervised, model agnostic, data-driven approach to Knowledge Extraction from heterogeneous KBs. We expect that the experimentation performed on Berenson's network of acquaintances will help determine the feasibility of Knowledge Extraction from highly coherent and focused SDLs. This would motivate a further experimentation of this methodology with a broader scope and less converging sources. Moreover, it would help define a quality assessment framework for the extracted information.

Bibliography

[1] Nichols, Stephen G. "Time to Change Our Thinking: Dismantling the Silo Model of Digital Scholarship." Ariadne, no. 58 (30 January 2009)http://www.ariadne.ac.uk/issue58/nichols/

[2] Dibowski, Henrik, et al. "Using Knowledge Graphs to Manage a Data Lake", 2021.

[3] Darmont, Jérôme, et al. "Data lakes for digital humanities." Proceedings of the 2nd International Conference on Digital Tools & Uses Congress, October 15, 2020, 1–4.

[4] Takagishi, Akira. "A Twentieth-Century Dream with a Twenty-First -Century Outlook: Yashiro Yukio, a Japanese Historian of Western Art, and His Conception of Institutions for the Study of East Asian Art," in Asian Art in the Twenty-First Century, Williamstown (Mass.): Sterling and Francine Clark Institute 2007, 138–48.

Commercial crowdsourcing in digital humanities: prospects and ethical issues

Suviranta, Rosa
rosa.suviranta@helsinki.fi
University of Helsinki, Finland

Hiippala, Tuomo
tuomo.hiippala@helsinki.fi
University of Helsinki, Finland

This presentation discusses key issues in using commercial crowdsourcing in digital humanities. Traditionally, digital humanities have engaged volunteers for tasks like digitising and organising information (Dunn and Hedges, 2013; Carletti et al., 2013). However, not all fields in digital humanities can benefit from volunteer-based crowdsourcing. I argue that commercial crowdsourcing is a viable alternative for fields that cannot attract volunteers, provided that crowdsourcing is used in an ethically sustainable way. To do so, I propose solutions to a range of ethical issues related to fair pay and hidden labour on commercial crowdsourcing platforms. I also discuss linguistic and epistemic asymmetries between task requesters and the global crowdsourced workers and argue for the need to develop crowdsourcing methods that balance the needs of both ethics and data quality. To this end, I draw on examples from an ongoing project that uses crowdsourcing to create multimodal corpora and show how a combination of pedagogically motivated training, paid exams and multimodal instructions can mitigate these issues.

Crowdsourcing is a participatory method in which an individual, an institution, organisation or company can request an undefined group of workers with varying knowledge and number to perform a task through an open call (Carletti et al., 2013: 1–2). Crowdsourcing can take place on commercial and non-commercial platforms, and the tasks can range from data labelling to content creation (Dunn and Hedges, 2013). How crowdsourcing is understood depends on the field of research. Computer vision, for example, uses crowdsourcing to create training data for algorithms by decomposing complex tasks into piece-meal work and distributing this effort among paid non-expert workers on online platforms (Kovashka et al., 2016).

In digital humanities, crowdsourcing is often associated with the galleries, libraries, archives and museums (GLAM)

domain. In this context, crowdsourcing is a way of engaging enthusiasts with intrinsic incentives to perform tasks for free and for the 'common good' (Daugavetis, 2021). Consequently, the ethos among digital humanities is to use crowdsourcing to engage volunteers to interact, explore and contribute to the research at hand (Dunn and Hedges, 2013; Terras, 2015).

However, not all fields under the umbrella of digital humanities can benefit from volunteer-based crowdsourcing. One such example is the emerging discipline of multimodality research, which studies the way human communication relies on intentional combinations of expressive resources. The discipline is currently undergoing a shift toward a more data-driven direction due to increased calls for validating theories of multimodality through empirical research (Pflaeging et al., 2021). This shift has also brought multimodality research in contact with digital humanities, especially within the paradigm of 'distant viewing' (Arnold & Tilton, 2019), which applies computational methods to large-scale analysis of visual materials (Hiippala and Bateman, 2021). Together with computational methods, commercial crowdsourcing has been identified as a potential way of increasing the size of corpora studied in multimodality research (Hiippala et al., 2019).

However, any use of commercial crowdsourcing must acknowledge the ethical issues and pitfalls related to crowdsourcing platforms. As a part of the novel platform economy, crowdsourcing lacks regulation which enables exploitative practices (Schmidt, 2017). Labour rights are largely absent, and the pay is usually far from a living wage, and there are no rules or stipulations for a minimum wage. Moreover, new workers often need to perform several months of non- or low-paid work in the form of qualification labour to access well-paid tasks (Kummerfeld, 2021). Qualification, or ranking, is a social reward technique to increase standing and reliability in the crowdsourcing community by completing tasks successfully (Dunn and Hedges, 2013: 152-154), which the requesters often use as a quality control tool to filter out low-performing workers (Kummerfeld, 2021: 343). Many workers also speak English – the lingua franca of crowdsourcing platforms – as a foreign language, which can lead to misunderstandings, rejected work and payment refusal.

Although the platforms enable exploitative practices, requesters can influence the conditions and for crowdsourced work. Firstly, the workers must be compensated appropriately and paid at least a minimum wage. Although ethically-sustainable crowdsourcing is not cheap, it is considerably cheaper than using experts (Hiippala et al., 2021: 673). Secondly, requesters can recruit workers with fewer qualifications, while maintaining quality by combining pedagogically-motivated training, paid exams and multimodal instructions. Pedagogically-motivated training allows the workers to learn the task through trial and error. Subsequently, paid exams filter the workers to perform the actual task. Pairing the training with the paid exam for selecting the workers ensures that even if a worker fails the exam, they are compensated for their effort. Finally, multimodal instructions, which combine text and illustrations, can support workers with limited language skills.

Bibliography

Arnold, T. and Tilton, L. (2019). Distant viewing: analyzing large visual corpora. *Digital Scholarship in the Humanities*, 34(supplement 1): i3–i16.

Carletti, L., Giannachi, G., Price, D., McAuley, D. and Benford, S. (2013). Digital humanities and crowdsourcing: An exploration. *Museums and the Web*. Available at: https://ore.exeter.ac.uk/repository/bitstream/handle/10871/17763/Digital%20Humanities%20and%20Crowdsourcing%20-%20An%20Exploration.pdf (Accessed: 21.4.2022).

Daugavietis, J. (2021). Motivation to engage in crowdsourcing: Towards the synthetic psychological–sociological model. *Digital Scholarship in the Humanities*, *36*(4): 858-870.

Dunn, S. & Hedges, M. (2013). Crowd-sourcing as a component of humanities research infrastructures. *International Journal of Humanities and Arts Computing*, 7(1-2): 147–169.

Hiippala, T., Alikhani, M., Haverinen, J., Kalliokoski, T., Logacheva, E., Orekhova, S., Tuomainen, A., Stone, M. and Bateman, J. A. (2021). AI2D-RST: A multimodal corpus of 1000 primary school science diagrams. *Language Resources and Evaluation,* 55(3): 661–688.

Hiippala, T. and Bateman, J. A. (2021). Semiotically-grounded distant view of diagrams: insights from two multimodal corpora. *Digital Scholarship in the Humanities*. Available at: 10.1093/llc/fqab063 (Accessed: 21.4.2022).

Kovashka, A., Russakovsky, O., Fei-Fei, L. and Grauman, K. (2016). Crowdsourcing in computer vision. *Foundations and Trends in Computer Graphics and Vision,* 10(3): 177–243.

Kummerfeld, J. K. (2021). Quantifying and avoiding unfair qualification labour in crowdsourcing. *In 'Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers),* Association for Computational Linguistics, Online, pp. 343–349. Available at: https://aclanthology.org/2021.acl-short.44 (Accessed: 21.4.2022).

Pflaeging, J., Wildfeuer, J. and Bateman, J. A. (eds.) (2021). *Empirical Multimodality Research: Methods, Applications, Implications* . Berlin and Boston: De Gruyter.

Schmidt, F. A. (2017). Digital labour markets in the platform economy. *Mapping the Political Challenges of Crowd Work and Gig Work, 7,* 2016.

Terras, M. (2015). Crowdsourcing in the digital humanities. In S. Schreibman, R. Siemens & J. Unsworth. (eds), *A New Companion to Digital Humanities*. Wiley, pp. 420–438.

Geographic analysis of published guidebooks and personal diaries on the diversity of city image in the Edo period

SUZUKI, Chikahiko
ch_suzuki@nii.ac.jp
ROIS-DS Center for Open Data in the Humanities, Japan; National Institute of Informatics, Japan

KITAMOTO, Asanobu
kitamoto@nii.ac.jp
ROIS-DS Center for Open Data in the Humanities, Japan; National Institute of Informatics, Japan

Introduction

In Japan, guidebooks became popular during the Edo period (17th to 19th century). The contents of the guidebooks were cited in various genres, such as literature, leading to the spread of a general *Edo* (江戸 present-day Tokyo) city image [1]. Even today, the public perception of Edo is based on these images. Many travel diaries written in the Edo period are preserved today, shedding light on personal activities in Edo city. Akio Suzuki enriched the image of Edo by comparing the general image in the guidebooks with the descriptions in some travel diaries [2]. An even broader image of Edo can be reproduced by quantitatively expanding Suzuki's method. Such an expanded method would be comparable to tourist information from today's online reviews.

For the aforementioned purpose, apart from increasing the amount of information from publications that form the basis of the city image, we organized and considered various individual activities written in diaries as historical activity records. We integrated the findings of the previous studies that compared some of these recorded activities. Therefore, we proposed a method for structuring the description of the diaries and guidebooks using identifiers. These identifiers helped connect place names that appeared in documents with modern-day geographic information.

Materials and Tools

As a record of individuals that corresponds to "online reviews," we selected the diary of three people with different attributes who visited the city of Edo during end of the Edo period [3]: *Hachiro Kiyokawa* (清河八郎), a *bushi* (武士, literally warrior or *samurai*); *Senkou-In* (泉光院), a priest of *Shugen-do* (修験道); and Heinrich Schliemann, a German archaeologist who visited Edo after the shogunate opened the country. First, we extracted the date (original Japanese calendar notation), converted Christian era data using HuTime [4], place names, and actions from the diaries of these three people. Second, we linked GeoLOD [5] identifiers to the extracted place names. For the comparison with the extracted description, we used a database of illustrations of guidebooks in the Edo period [6].

GeoLOD is a web service for the registration, management, display, and sharing of geographic information by assigning identifiers (GeoLOD ID) to place names. With GeoLOD, we linked place names appearing in multiple sources with their GeoLOD ID's latitude and longitude. GeoLOD enables the integration of not only the place names but also various information contained in the original material into the "geospace", or information space based on geographic information. We then analyze this information on the geospace.

Results and Discussion

Figure 1 visualizes the activities of *Hachiro Kiyokawa* in Edo, achieved by linking geographic information with his narrated activities [7]. Furthermore, GeoLOD allowed us to not only link activities and geographic information but also to integrate and compare diary descriptions and information in publications.

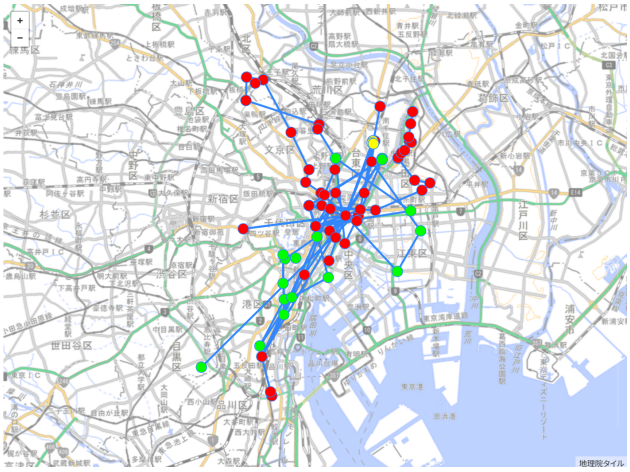


Fig.1

Linking Hachiro Kiyokawa*'s movement with a modern-day map*

By comparing descriptions, we found that all three people with different attributes visited the most famous places of the time, such as *Atagoyama* (愛宕山) and *Sensoji* Temple (浅草寺). Furthermore, by linking materials using GeoLOD, we could compare the diary descriptions with the illustrations of famous places in guidebooks. Thus, we could clarify the difference between the situation of the famous places that each traveler paid attention to and the general image depicted in the publication. We noted various situations in Edo by finding the differences between the basic stereotyped images and individual observations (Fig. 2). Furthermore, by using identifiers, we could easily add and compare information.

Enfin nous arrivâmes au grand Propylaeum du Temple d'Asaksa-Quannon, d'où une longue et belle avenue conduit à la porte intérieure du Temple. Cette avenue est bordée des deux côtés d'un immense nombre de boutiques qui forment un vaste bazar, où on vend principalement des jouets d'en fants, des idoles, des ornements de femmes et surtout des épingles à cheveux dorées, à têtes de verre, creuses et remplies d'un liquide coloréet de paillettes d'or. Cette avenue était encom brée de femmes, d'enfants, de fainéants, d'acheteurs et de vendeurs. (from *La Chine et le Japon au temps present* P.147)

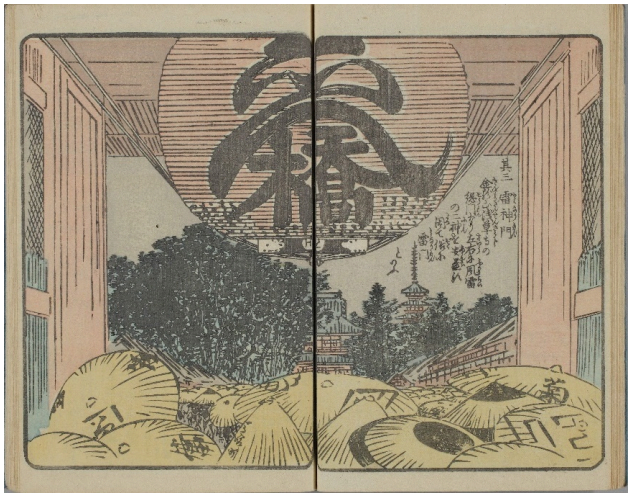


Fig. 2
Schliemann's description of Sensoji *Temple in an original French version (above) and the illustration of a Japanese guidebook of the same period (bottom). The products of the stalls are described in detail, reinforcing the general image. Meanwhile, no description is given of the large lantern that represents* Sensoji *Temple.*

By developing connections with geographic information, we could collect data on the movements and behaviors of people described in historical materials as historical activity records. Future work may link various data extracted from historical materials, such as historical activity records, with spatiotemporal information.

We further utilized GeoLOD to publish "edomi" [8] by linking the databases with geographic information. "edomi" is a data portal site that offers a panoramic view of the city of Edo and various information on the Edo period. Users can search and utilize historical information across categories based on the current perspective (Fig. 3). By linking the movement of the three people with edomi, we plan, in a future work, to not only present geographic information but also clarify the relation between tourists and famous places in Edo.



Fig. 3
Category list of content available on edomi: travel, shopping, map, gourmet, politics, economy, disaster, and learning (Japanese only as of November 2021)

Bibliography

Ichiko, N. and Suzuki, K. (2009). Edo-Meisho-Zue wo Yomutameni, *Shintei Edo-Meisho-Zue Bekkan 2*, Chikuma Shobou (in Japanese)

Suzuki, A. (2001). *Edo no Meisho to Toshibunka,* Yoshikawakoubunkan (in Japanese)

Kiyokawa, H. (1993). *Sai-Yu-Sou*, Iwanami Shoten (in Japanese). Ishikawa, E. (1994). *Senko-In Edo Tabi Nikki*, Kodansha (in Japanese). Schliemann, H. (1867). *La Chine et le Japon au temps present*, Libr.Centrale

HuTime. http://www.hutime.org/

GeoLOD. https://geolod.ex.nii.ac.jp/

Suzuki, C. and Kitamoto, A. (2020). Creating Structured and Reusable Data for Tourism and Commerce Images of Edo: Using IIIF Curation Platform to Extract Information from Historical Materials, *Digital Humanities 2020* (in Japanese)

Kitamoto, A., Suzuki, C., Terao, S., Horii, M. and Horii, H. (2020). Construction of Edo Big Data Based on Structuring and Integrating Historical Placenames on Spatial Historical Sources, *Jimmoncon2020*, pp. 171-178

edomi. http://codh.rois.ac.jp/edomi/

Digital Humanities in the Wild: Bringing Humanistic Pedagogy to Open Source

Tagliaferri, Lisa
lisa.tagliaferri@gmail.com
Rutgers University

What do the digital humanities look like outside of the library and the classroom? This paper presents a possible site of digital humanities pedagogy within the realm of open source.

In 2021, myself and a team built out and began to populate Sourcegraph Learn (https://github.com/sourcegraph/learn), an open source and open access web project that provides educational resources for software developers. A project with only a few contributors on a singular team is not necessarily in the spirit of open source, so during the month of October the team worked with the momentum of Hacktoberfest, the yearly celebration of open source, to invite and encourage more contributors to become involved in our project. We created a Hacktoberfest-specific readme.md file with instructions, added relevant tags, and created issues specific to programming languages and error messages.

Software developers were invited to contribute troubleshooting guides to the project. These guides would go into depth regarding common error messages that programmers encounter. For example, titles of these technical tutorials included "How to troubleshoot Java ArrayIndexOutOfBoundsException," or "How to troubleshoot Python AttributeError." Each tutorial included the full text of the error message, a way to replicate the error message, and two or three ways that a developer could recover from the error message. The troubleshooting guides are all available under the "troubleshooting" tag of the Sourcegraph Learn site (https://learn.sourcegraph.com/tags/troubleshooting).

Altogether, over 40 Git issues were created soliciting contributions, nearly 40 pull requests were opened by prospective contributors, and over 30 high-quality pull requests were merged from contributors across 5 continents. These community-based efforts represent a global, distributed outcome of employing pedagogical techniques such as student-centered learning through encouraging practitioners to share what they have learned with others.

This paper will present the approach to documentation, templatizing Git issues, and collaborative code reviews in order to demonstrate how others may be able to set up open-source humanities computing projects to encourage public contributors. In particular, I will draw from the body of work related to crowdsourcing in relation to museums and archives, including Mia Ridge's edited collection *Crowdsourcing our Cultural Heritage* (2014), Melissa Terras's "Crowdsourcing in the Digital Humanities" (2015), and the Getty Museum's and Folger Library's respective Zooniverse projects to encourage crowdsourcing. Additionally, this paper will explore how engaging in this space may benefit both open source and the humanities.

Following the presentation, the author would be interested in hearing from other researchers who may have accepted open-source contributors, and how others may have scaled their approach to digital pedagogy through open source.

Bibliography

Deines, N. (2018). "Six Lessons Learned from Our First Crowdsourcing Project in the Digital Humanities." *Getty Iris Blog*, https://blogs.getty.edu/iris/six-lessons-learned-from-our-first-crowdsourcing-project-in-the-digital-humanities/.

Folger Library. *Shakespeare's World*. https://www.zooniverse.org/projects/zooniverse/shakespeares-world.

Ridge, M. (2014). *Crowdsourcing Our Cultural Heritage*. Ashgate.

Terras, M. (2016). "Crowdsourcing in the Digital Humanities." *A New Companion to Digital Humanities*, edited by Schreibman, Siemens, Unsworth, Wiley-Blackwell, pp. 420-439.

Integrating the Japanese Archaeological Dataset into the ARIADNEplus Data Infrastructure

Takata, Yuichi
takata-y23@nich.go.jp
Nara National Research Institute for Cultural Properties, Japan

Yanase, Peter
yanase-p7g@nich.go.jp
Nara National Research Institute for Cultural Properties, Japan

Niccolucci, Franco
franco.niccolucci@gmail.com
PIN, Italy

The Comprehensive Database of Archaeological Site Reports in Japan (SORAN) is Japan's largest repository and aggregator of archaeological data and information. On November 25, 2021, it contained 29,861 full-text PDF copies of fieldwork reports, 112,487 pieces of bibliographical information, and 138,552 sets of detailed metadata of archaeological interventions (Comprehensive database of archaeological site reports Japan, 2015). It is an immensely popular service that in 2020 had over 13.5 million visits and 78.5 million page views. However, because the service was originally built to satisfy the domestic market, its spatial coverage is delimited by national borders and its user base by a language barrier. To overcome these hurdles, the service's operator, the Nara National Research Institute for Cultural Properties (NABUNKEN), decided to integrate a significant part of SORAN's data into the data infrastructure managed by the Archaeological Research Infrastructure for Archaeological Data Networking in Europe (ARIADNEplus), a project whose original goal was "to provide open access to Europe's archaeological heritage and overcome the fragmentation of digital repositories, placed in different countries and compiled in different languages" (Niccolucci and Richards, 2019: 7).

The most readily visible part of ARIADNEplus is the ARIADNE Portal, a website providing access to the ARIADNE Catalogue containing the aggregated metadata of the project partners. The Portal is a tool enabling both cross-border/cross-institution resource discovery and data manipulation. This, in practicality, means that after integration, SORAN's data will be part of an extensive European dataset searchable and processable via a common user interface.

The ARIADNE Catalogue is searchable according to the three facets of "when" (time), "where" (space), "what" (object), as well as keywords drawn from controlled vocabularies. While SORAN itself supports information retrieval in a similar manner, the way relevant information is implemented and presented is radically different from ARIADNEplus. Therefore, NABUNKEN and ARIADNEplus had to collaborate closely in a long integration process involving data cleansing, schema transforming, and concept mapping.

Mapping SORAN's internal data schema to ARIADNE's ontology was a largely technical step. Although the two schemas are different in concept and file format, the mapping could be done in a few weeks. Mapping the Japanese data to the facets of "When?," "Where?," "What?" was more complicated. The first facet required spatial coordinates to be converted to comply with the WGS84 (World Geodetic System 1984), which a significant amount of the original data did not follow. The second facet required temporal information to be linked to definitions stored on PeriodO (a multilingual gazetteer of temporal information) (PeriodO, no date). In Japan, the exact temporal limits of historical periods are often debated and difficult to define. Thus we had to enlist an interdisciplinary team of experts to assist us with that. The final facet of objects required the most work as it involved mapping culture and discipline-bound terms to the Getty Art & Architecture Thesaurus.

An important aspect of the mapping process of these facets was that we had established several re-usable rules for transforming and mapping Japanese information into intelligible English during the process. From the very start, the mapping project intended to make both the process and choices transparent and develop methods that other institutions could re-use and adopt in similar situations.

Collaboration with ARIADNEplus is more than just providing a dataset for an international data infrastructure: it involves a lot of discussions between the various partner institutions involved. Each member can gain new insights by learning about both international best practices and local solutions. This not only helps foster interoperability but lowers development costs as well.

Japan is the first Asian country to integrate its data in ARIADNEplus, but hopefully not the last. Our presentation aims both to explain the possibilities presented by this international collaboration and showcase our solutions developed in the process.

Bibliography

Comprehensive database of archaeological site reports Japan (2015). https://sitereports.nabunken.go.jp/en/ (Accessed 25 November 2021).

Niccolucci, F. and Richards, J. (2019). ARIADNE and ARIADNEplus. In Niccolucci, F. and Richards, J. (eds), The ARIADNE Impact. Budapest: Archaeolingua Foundation, pp. 7-25. https://doi.org/10.5281/zenodo.3476711

PeriodO – Periods, Organized (no date). https://perio.do/en/ (Accessed November 25, 2021.)

Queer Coding the Audio Archive: Linked Data and the Lesbian Organization of Toronto (LOOT) Oral History Tapes

Tayler, Felicity
felcity.tayler@uottawa.ca
University of Ottawa, Canada

Crompton, Constance
constance.crompton@gmail.com

University of Ottawa, Canada

History meets material meets digital: the digital and analogue materialities of oral history interviews from the archival Fonds of the Lesbian Organization of Toronto (LOOT), prompt us to think through the ethics of representation in digital imaginaries. Our short paper on Humanities-centred linked data and study of literary audio in Canada offers a case study of digitization and metadata description as a contribution to the larger SSHRC-funded Spoken Web partnership, a partnership dedicated to the study of Canadian literary audio archives. The LOOT audio offers significant challenges for digitization, analogue and digital preservation, archival metadata description and ethical linked data creation. As Becki Ross has observed in *The House That Jill Built: A Lesbian Nation in Formation,* this oral history audio reflects a particular political context and community milieu in Toronto, Canada's largest city. We hypothosize that when linked to the Spoken Web corpus of literary audio data, it will make visible hidden elements of poetic community: lesbian identity, women's labour, feminist media, leftist political organizing, queer social spaces, and the print and audio cultures of racialized women.

Our approach treats the semantic web and its code as a representative medium, and reflects on power of CIDOC-CRM, Schema.org, and BIBFRAME to represent this nuance. This awareness of code as language with representational power is key when listening for queer coding in the digitized audio archive, and the tensions between making sure that gay lived experience is represented on the semantic web on the one hand, and the representational limitations of the ontologies that underpin the semantic web one the other (Hedley and Janzen Kooistra).

We align our linked data work with a feminist practice of collection, curation, and visualization of data as a counter-discourse to the networks of communicative capitalism (Dean). Even amongst critiques of co-optation, intersectional feminisms are "central to the identity and the methodologies of the digital humanities as a field" (Losh and Wernimont, xi). Catherine D'Ignazio and Lauren Klein propose "feminist data visualization" as an active practice to make explicit the ways in which data collection and display contributes to the marginalization of voices, economic or social positions; generates emotional bonds through qualitative arguments; or conceals underlying power dynamics that mark the provenance of data sets – and yet, that visualization is dependant on the representational power of ontologies we choose to to structure the data. We also look to the work of Safiya Umoja Noble, Ruha Benjamin, and Jennifer Wemigwans to ask: can these semantic linked networks be perceived as hopeful social movements, or do they perpetuate data discrimination through computational algorithms of oppression? Does our metadata reproduce the social inequities embodied in these oral history interviews? Is listening with care to oral histories a way to move slower and get closer to respectful engagement with the humans behind the data? And how are the absences, what you cannot see, just as important as what is made visible by linking up this data across archival collections?

Bibliography

Benjamin, Ruha. *Race after Technology : Abolitionist Tools for the New Jim Code*. POLITY PRESS, 2019.

Dean, Jodi. *Democracy and Other Neoliberal Fantasies Communicative Capitalism and Left Politics.* Duke University Press, 2009.

Hedley, Alison, and Lorraine Janzen Kooistra. "Prototyping Personography for The Yellow Nineties Online." *Bodies of Information: Intersectional Feminism and the Digital Humanities*, edited by Elizabeth Losh and Jacqueline Wernimont, University of Minnesota Press, 2019, pp. 157–72.

Losh, Elizabeth, and Jacqueline Wernimont, *Bodies of Information: Intersectional Feminism and the Digital Humanities*. University of Minnesota Press, 2019

Noble, Safiya Umoja. *Algorithms of Oppression : How Search Engines Reinforce Racism*. New York University Press, 2018.

Ross, Becki L. *The House That Jill Built : a Lesbian Nation in Formation*. University of Toronto Press, 1995.

Wemigwans, Jennifer. *A Digital Bundle: Protecting and Promoting Indigenous Knowledge Online*. University of Regina Press, 2018.

The Japanese Small World of Words. Investigating meaning through a large-scale crowdsourcing study of word associations.

Telegina, Maria
maria.telegina@mail.u-tokyo.ac.jp
University of Tokyo, Japan

De Deyne, Simon
simon.dedeyne@unimelb.edu.au
University of Melbourne, Australia

Joyce, Terry
terry@tama.ac.jp
Tama University, Japan

Miyao, Yusuke
yusuke@is.s.u-tokyo.ac.jp
University of Tokyo, Japan

Language is the bridge that connects cultures, but knowing whether a foreign language word has the same meaning and context as in their mother tongue is one of the most challenging tasks for foreign language learners and educators. Previous studies (e.g. De Deyne, Verheyen and Storms, 2016) suggest that word association data contains semantic, cultural, and extra-linguistic information that underlies such common knowledge providing a unique method to compare meaning across cultures.

In contemporary linguistics and psychology, word association data have been rediscovered as a source of information for research on language and the mind. Word associations allow us to investigate a wide range of phenomena, including demographic-dependent differences in language use (Garimella et al., 2017), lexical centrality and semantic similarity (De Deyne et al., 2019), language development, and age-dependent changes in concept (Wulff *et al.*, 2019).

There are three previously collected data sets of the Japanese word associations: Japanese word association norms collected by Umemoto (1969), the Associative Concept Dictionary (Okamoto & Ishizaki, 2001), and the Japanese Word Association Database (JWAD; Joyce, 2005). The word association norms by Umemoto the stimuli set consisted of 210 words and responses were collected from 1000 respondents. The Associative Concept Dictionary consists of two data sets: one has 1656 stimuli, approximately 130,000 responses in total; another one has 1055 stimuli and approximately 250,000 responses. JWAD consists of 104,800 associative responses to 2099 stimuli. In all three cases, the volume of the data, demographics of the respondents, and information on word relations were limited due to the challenge of collecting data at scale. Moreover, methodological differences such as instructions complicate the comparison across languages that use slightly different procedures.

This paper presents a project aiming to create a large-scale Japanese associative database as part of the multilingual Small World of Words project (SWOW-JP). This project uses online crowdsourcing as the primary data collection. Information about the project is distributed via social media, and the word association collection is organized via the project's web page. Crowdsourcing has allowed us to overcome volume and demographic limitations, already resulting in a dataset covering over 165,000 responses and more than 4000 participants. The average age of all participants was 37 (SD = 15 years). Participants were represented across all prefectures, with the top 3 coming from Tokyo (17%), Kanagawa (10%), and Aichi (5%). In a follow-up study, Japanese native-speaking respondents will be verifying relations between words via a citizen-science platform.

Japanese is one of 18 languages currently included as part of the international collaborative Small World of Words project. Datasets in several major world languages are now available (Dutch, +18,000 cues, English, +14,000 cues) or prepared for publication (Spanish +13,000 cues, Mandarin, +10,000 cues). The English and Dutch databases have already been downloaded by more than 3500 researchers. The simultaneous collection and comparable methods used across languages, including Asian languages such as Cantonese, Mandarin, and Korean, provide a unique resource for comparative analyses. Besides new psycholinguistic resources and demographic-aware semantic representations in Japanese, the SWOW-JP project will benefit from collected data in other languages, including logographic languages and the scientific lingua franca (English). We expect this to be instrumental in addressing theoretical questions about conceptual universality, providing a benchmark for NLP models, and supporting several applications, such as bilingual vocabulary learning.

The specifics of the Japanese writing system will also provide new opportunities and challenges compared to existing Indo-European datasets. One possibility is that the use of word forms across multiple writing systems might elicit different semantic representations (e.g., a word in hiragana vs kanji). The global network structure in Japanese semantic networks might be considerably different from other languages. This would be supported by previous findings in a cross-linguistic comparison of word associations across 12 languages (Miron and Wolfe, 1964), showing that Japanese word associations are the most stereotypical (highest agreement among responses). Altogether, we believe the SWOW-JP project has the potential to revisit old and new questions systematically and comprehensively across many disciplines. Furthermore, beyond providing a source of data for linguists and psychologists, Japanese is a language spoken by over 120 million speakers and taught in over 136 countries, which opens several exciting avenues for multilingual comparative research, foreign language education, and other disciplines investigating the interaction between language thought and culture.

Metadata, Data, and Datasets: An Exercise in Excising the Web Archive for Public Consumption

Thomas, Grace
grth@loc.gov
Library of Congress, United States of America

Dooley, Chase
cdool@loc.gov
Library of Congress, United States of America

Innovation is sparked in the absence of access to a service or resource. Or, in the case of libraries, merely access itself. As a core value of librarianship, access stands paramount to the mission of libraries and, therefore, so does the necessity to innovate in the face of a gap in access. The Web Archiving Team at the Library of Congress found ourselves facing such a gap when we stood on the edge of our digital cliff and peered out over it and into the three Petabytes (PB) of web archives swimming in our lake of data. In this talk we introduce the ways in which we bridged that gap by publicly providing basic but rich technical metadata about web resources in the form of crawl indexes (CDX's). First, we will briefly describe the methodology behind the Library's harvesting practices and the difficulties they have made in presenting data in bulk to researchers. Then, we will touch on the technical aspects: our approach, research, tools used, and results. Finally, we will discuss the impact this work has on the digital humanities community, and invite researchers to experiment for themselves.

The Library of Congress web archives are organized among 80 thematic and event-based collections, and contain websites representing a broad range of subjects, languages, file formats, and topic areas, with a mix of crawling and access permissions based on the country of publication and type of website. Areas such as: government; non-profit and for profit organizations; journalism and news; and creative sites are collected from the United States and throughout the world. Active Library subject specialists maintain and continually refine collections such as the Indian Political and Social Issues Web Archive collection which has been collecting content in English, Hindi, Urdu, Marathi, and Bengali for two years, and the United States Elections Web Archive collection, which has been collecting campaign material in English and Spanish during every national election season since 2000.

Although the archive's contents are nominated, cataloged, and made available on loc.gov according to the event and thematic collections, the harvesting takes a different form. Crawls are performed weekly, monthly, and quarterly, and each crawl acts as a bucket for any seeds in the collections set to crawl at that particular frequency. Harvesting in this ‚bucket-style' allows the Web Archiving Team to streamline many aspects of crawling over 14,000 seeds (or ‚websites') at any given time and allows harvesting to reflect the specific website's frequency of change.

This also means that the container files (Web ARChive or ‚WARC' files) holding the harvested web objects were collected ‚bucket-style' and may contain objects from 100 different websites, representing 20 different collections with varying access permissions in a single container file. Hence the issue with providing bulk data to researchers. Even after the Web Archiving Team received access to cloud services in 2018, and could experiment with access at scale, WARCs still could not be presented to researchers as-received because of the mixed permissions. We looked to existing tools and organizations for inspiration to find a path that would work for us. We ultimately adapted work from Common Crawl's public Github repository (https://github.com/commoncrawl/cc-pyspark, https://github.com/commoncrawl/cc-mrjob). We also took cues from Archives Unleashed (Ruest et al., 2021) and ArchiveSpark (Holzmann et al., 2016) by utilizing EMR (Elastic Map Reduce) and Spark to process the CDX's.

The high-level process is very simple: transform the CDX's into a format that is better suited for EMR and data processing, then query to reduce down to the desired output. The idea is modeled after Common Crawl's approach in which they transform their CDX's into Parquet format for enhanced compression and easier consummation into DataFrame objects that are part and parcel of Spark and other such Big Data processes.

The ability to cleanly excise a part of the archive represented by its metadata is a huge step forward in providing bulk access to the archive. After making the first datasets publicly available (https://labs.loc.gov/work/experiments/webarchive-datasets/) we have seen use by students, information professionals, and researchers. Researchers use the metadata to load specific archive captures from the Library's Wayback Machine instance, and ‚rehydrate' the text by scraping the archived captures to create a text-based dataset of their own. While this method requires some technical skill and processing power, we see countless opportunities to create web archive-based datasets for any discipline, given the Library's broad collecting range, and look forward to fielding researcher requests.

Bibliography

Holzmann, H., Goel, V., Anand, A. (2016). ArchiveSpark: Efficient Web Archive Access, Extraction and Derivation. *JCDL '16: Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*. Newark, NJ, pp. 83-92.

Ruest, N., Fritz, S., Deschamps, R. et al. (2021). From archive to analysis: accessing web archives at scale through a cloud-based interface. *International Journal of Digital Humanities*, 2: 5-24.

Infrastructural Sovereignty in the Black Atlantic

Thorat, Dhanashree

dhanashreepune@gmail.com
Mississippi State University, United States of America

In June 2019, Google announced a new undersea fiber-optic cable line connecting Portugal and South Africa and named the line "Equiano" after the eighteenth-century Black man who was sold into the Transatlantic Slave Trade and later purchased his own freedom. Undersea fiber-optic cable lines are the material backbone of the global Internet network and responsible for almost all Internet connectivity worldwide. In this presentation, I read Google's infrastructural politics against the grain of Equiano's autobiographical account, *The Interesting Narrative of the Life of Olaudah Equiano, or Gustavus Vassa, the African* (1789), to locate how the afterlives of slavery and colonialism manifest in Internet infrastructure. Google's choice of name for the new cable line enmeshes a historical moment with contemporary infrastructure threaded through by the iterative logic of racial subjugation. I argue that Internet infrastructural projects like Google's new line reroute colonial ambitions, operating in the oceanic pathways of the slave trade and on the ideologies of racial capitalism now trading in datafied Black bodies. In the context of Google's project, I pursue the deeper implications of Equiano's invocation to ask how Internet corporations might acknowledge their own complicity in inherently inequitable projects of techno-modernity.

Methodologically, I draw on literary criticism and rhetorical analysis to illuminate historically situated (infra)structural modalities by specifically focusing on Equiano's autobiography and press releases by Google. I build on extant scholarship in Black digital humanities (Jessica Marie Johnson, 2018; Kim Gallon, 2016) and postcolonial and decolonial approaches to data justice (Paola Ricaurte, 2019, Syed Mustafa Ali, 2016; Roopika Risam, 2018; Abeha Birhane, 2019; Marisa Duarte, 2017; Noopur Raval, 2019) to examine how contemporary infrastructural projects undertaken by Internet era companies like Google, Microsoft, and Facebook ultimately enable datafication projects and algorithmic violence. I advance two threads of conversation: First, my paper highlights the implications of such cable projects for infrastructural sovereignty when Western and capitalist corporations drive the development of Internet infrastructure in postcolonial nation-states like Nigeria. Second, I forward Equiano's testimony and his unrestrained desire for freedom as a call to reimagine infrastructural politics and data justice.

As Google begins construction on this new cable project, other tech corporations like Facebook have already made parallel competitive moves in other postcolonial locations. This model of Western ownership of key communications infrastructure undermines the long arc of independence struggles which wrested control away from colonial empires so that postcolonial countries could own and govern their own critical infrastructures. Until corporations and governments reconcile the violent histories that Internet infrastructures are part of and call into being, these projects will continue to channel reductive ideologies about the Global South and undermine the infrastructural sovereignty of postcolonial nations. The need for robust Internet infrastructures is likely to continue increasing in postcolonial countries with increases in data traffic, AI technologies, and ubiquitous computing worldwide, and it is imperative that we take account of and divest Internet infrastructures from the racial genealogies they emerge from.

Bibliography

Ali SM. (2016). "A brief introduction to decolonial computing." *XRDS: Crossroads*. 22 (4):16-21.

Birhane A. (2019). "The Algorithmic Colonization of Africa." *Real Life Magazine*. 18 July, 2019.

Equiano O. (1789). *The Interesting Narrative of the Life of Olaudah Equiano, or Gustavus Vassa, the African*. Vol I & II London. Accessed on Documenting the American South (Docsouth) website.

Duarte M. (2017). *Network Sovereignty: Building the Internet Across Indian Country*. Seattle: U Washington Press.

Gallon K. (2016). "Making a Case for the Black Digital Humanities." *Debates in the Digital Humanities*. Minneapolis: U of Minnesota Press.

Johnson JM. (2018). Markup Bodies: Black [Life] Studies and Slavery [Death] Studies at the Digital Crossroads. *Social Text*. 36.4: 57-79.

Raval N. (2019). "An Agenda for Decolonizing Data Science." *Spheres: Journal for Digital Cultures*. 5.

Ricaurte P. (2019). Data Epistemologies, Coloniality of Power, and Resistance. *Television & New Media*. 20.4: 350-365.

Risam R. (2018). Decolonizing the Digital Humanities in Theory and Practice. *The Routledge Companion to Media Studies and Digital Humanities*. Ed. Jentery Sayers. New York: Routledge. Print.

Visualizing Rare Books: A Report on Manicule

Trettien, Whitney
trettien@english.upenn.edu
Price Lab for Digital Humanities

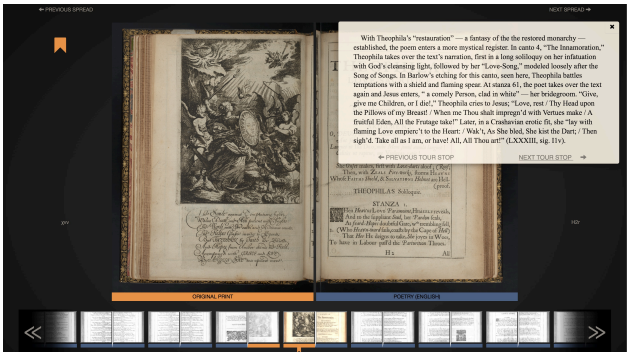
This short paper reports on Manicule, a web-based application for building digital tours of unique books. The application also includes a means of visualizing the structure of a codex, using the VisColl standard developed at the Schoenberg Institute for Manuscript Studies and implemented in, most recently, the BiblioPhilly database.

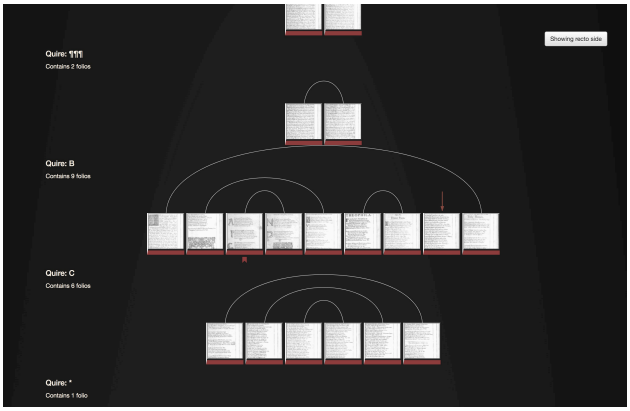
Since the New Bibliography of the first half of the twentieth century, bibliographic methods have been tethered to the needs of editing – namely, to identify from amongst a work's many reproductions the kernel of a single, reproducible edition. Of course, many textual theories have challenged the notion that any work can be reproduced as a single entity, to the extent that all texts are now seen as plural, polyphonic, variant, or marked by *mouvance* (Bornstein et al., 1993; Deegan et al., 2009; Egan, 2010; Pierazzo, 2015). This shift in thinking is reflected in the development of digital editing tools like Edition Visualization Technology, which allows editors to layer the digital text with annotations and cross-references, the Variorum Viewer of Frankenstein, or even the collation features of Mirador. Yet the idea that the goal of editing is to produce some form of text for reading and critical study remains baked into even the most variant-sensitive digital tools.

Meanwhile, book history – a more sociological offshoot of bibliography that has been growing steadily since the 1980s – has, contrary to bibliography and textual studies, begun to focus more on copy-specific work. For instance, book historians today increasingly pursue case studies of individual, idiosyncratic, and heavily "used books," to borrow Bill Sherman's term (Sherman, 2009). The scholarly telos of such work is very different from that in textual studies: whereas the latter tends to examine unique books only for what they might tell us about a literary work and its contemporaneous reception, the former imagines what such books can teach us about the long history of collecting and conservation (an interest that book history shares with digital humanities).

As a result of this split in focus, the digital editing tools devised under the regime of New Bibliography are not well suited for book historians working digitally with used books. For instance, a book historian does not often want to work with a text; thus TEI and TEI-based publishing technologies are useless to her. Rather, she is more likely to be interested in the text's material instantiation: how a codex has been formed and reformed over time as it passes through many different hands and institutions. Manicule is devised to meet these emerging needs in digital book history. Built by Liza Daly and Whitney Trettien, it is a standalone React/Redux application for presenting unique printed books and manuscripts in digital facsimile. It allows editors to:

- build guided tours through a book's distinguishing features,
- annotate the edges of interesting pages with extra information,
- categorize and color-code each page in the facsimile, giving a birds-eye view of a book's main elements,
- and visualize the book's structure.





Right now, the application is focused on visualizing and comparing Western codices. As part of our short paper, we hope to solicit feedback from ACH's international membership on how to build out functionality to fit other book forms, like tekagami, often folded like an accordion, or Arabic and other manuscript traditions that read from right to left. If we think about the work of books as assembling, as gathering together information into new constellations and configurations – whether made up of stitched paper quires or bundles of digital text and metadata – then we should have to-hand bibliographic tools that help see how they do this work. Manicule is a step toward building that toolkit.

Bibliography

Bornstein, G. et al., eds. (1993). *Palimpsest: Editorial Theory in the Humanities.* Ann Arbor: University of Michigan Press, 1993.

Deegan, M. et al., eds. (2009). *Text Editing, Print and the Digital World.* Burlington: Ashgate, 2009.

Egan, G. ed. (2010). *Electronic Publishing: Politics and Pragmatics.* Tempe: ACMRS, 2010.

Pierazzo, E. (2015 *). Digital Scholarly Editing: Theories, Models and Methods*. Burlington: Ashgate, 2015.

Sherman, W. (2008). *Used Books: Marking Readers in Renaissance England.* Philadelphia: University of Pennsylvania Press, 2008.

"Sweete Flowers and Odoriferous Beds of Spice": Sensory Mining Techniques to Trace Olfactory Orientalism

Tullett, William
william.tullett@aru.ac.uk
Anglia Ruskin University , Cambridge, United Kingdom

Menini, Stefano
menini@fbk.eu
Fondazione Bruno Kessler, Trento, Italy

Leemans, Inger
inger.leemans@huc.knaw.nl
KNAW Humanities Cluster, Amsterdam, Netherlands

How to compute the nose, mapping smells and odor perception? Olfactory informatics is a fast growing field, which aims to develop machine learning technologies to capture the enigma of olfaction. One challenge is the classification of flavors and fragrances, grouping odors into similar classes. For this task most projects work with chemical and physio-psychological data, and with the taxonomies which are in use in perfumery and odor industries.

The EU Horizon2020 *Odeuropa* project takes another turn, as it works with historical text and image collections from European digital heritage collections. Here, we find rich data on *historical scent perceptions* and *scent classifications*. How did people, both expert 'noses' and laymen, describe and classify scents in the period between 1600 and 1900? Here we present the first results of the NLP phase of the Odeuropa project, focusing on one specifically interesting 'class' of scents: 'oriental'.

Since the early twentieth century "oriental " has been a term used in perfumery to describe scents that include notes such as amber, sandalwood, and gum resins. Guerlain's *Shalimar* (1921) was one of the first 'oriental' fragrances. Today, many argue that the term is offensive and campaigns to remove the category from the perfumer's lexicon have proliferated. The use of the term in perfumery is the olfactory equivalent of the nineteenth and early twentieth-century literary and visual culture that, as Edward Said demonstrated, constructed the 'east' or 'the Orient' as sensual, primitive, irrational, and therefore opposed to the 'Occidental' West. However, recent scholarship has pushed the development of this orientalist discourse back into the early modern period.

In this paper we therefore seek to historicise the relationship between smell and 'the Orient' or 'the East' in collections of texts from the period 1600-1920: what did European noses across this period smell as 'Oriental' and what scents were associated with 'the Orient' or the 'East'? How did this change and how does this relate to the emergence of the term 'Oriental' in perfumery in the early twentieth century?

For this case study we analyze three corpora: EEBO (covering the timespan up to 1700), Project Gutenberg and Wikisource (using, in both, English documents up to 1920). In order to extract the potentially useful passages in the texts we defined a list of terms related to the Far East (e.g. *Orient*, *oriental*, *Indies*, etc), and a list of seed terms relating to smell (e.g. *smell*, *odor*, *scent*, *perfume*, *etc*.). From the corpora we extracted the passages containing at least one word affine to the 'oriental' topic and at the same time a term related to smell. We extracted then the nouns and adjectives close to the smell related words (limiting the range to 5 words before and after to reduce data noise), that may refer and describe the smells and the sources of smell related to the Orient, and that, associated with the year of

the document, can provide insights into how smells related the Orient change over the time.

For example, it is clear that in the early 1600s many of the key scents that were associated with the 'oriental' perfume category from the 1920s onwards were already part of how Europeans imagined the olfactory Orient, including amber, musk, rose, civet, and general references to 'spice'. These scents remained common throughout the time period. In the late 1700s jasmine and resins and in the first half of the nineteenth century woods, spices, and opium were then added into the imagined bouquet. This parallels the emergence of orientalising stereotypes in French cosmetic and perfume advertising during the late eighteenth and early nineteenth century. The spaces and feelings evoked by these scents also changed: earlier descriptions emphasize fragrant, odoriferous, and sweet scents that are linked to tales about fragrant Arabian coasts or biblical incense. However, beginning in the seventeenth century the Oriental associations with sensuality and eroticism that 'Oriental' perfume advertising would play on - most famously in the case of Yves Saint Laurent's 'Opium' - clearly appear as terms such as 'ravishing' and later 'intoxicating' become prominent in the data. This matches the importance of the harem in literary and visual evocations of the Orient - as in Delacroix's painting of the *"Women of Algiers>"*, which was said by Baulelaire to "exhale the heady scent of a house of ill repute".

In this contribution we present our results and discuss what conclusions we can derive from this exploration. We aim to show that both the smells linked to the 'Orient' and the feelings, atmosphere, and spaces they evoked had a long history before it consolidated as a perfume group, but that the population of this class and its evaluation changed over time. Our plea is that Digital Humanities technologies for mining scent perception can bring a new perspective to the field of odor informatics.

Bibliography

Candau, J. (2016). L'anthropologie des odeurs : un état des lieux. *Bulletin d'études orientales*, no. LXIV: 43–61.

Candau, J, and Wathelet, O. (2011). Les catégories d'odeurs en sont-elles vraiment ?. *Langages*, no. 181: 37–52.

Classe, C. (1992). The Odor of the Other: Olfactory Symbolism and Cultural Categories. *Ethos* 20, no. 2: 133–66.

Huang, X. (2016). Deodorizing China: Odour, Ordure, and Colonial (Dis)Order in Shanghai, 1840s–1940s*. *Modern Asian Studies* 50, no. 3: 1092–1122.

Kettler, A. (2020). The Smell of Slavery. Olfactory Racism and the Atlantic World. Cambridge: Cambridge University Press, 2020.

Lisena, P. and Van Erp, M. and Bembibre, C. and Leemans, I. (2021). Data Mining and Knowledge Graphs as a Backbone for Advanced Olfactory Experiences. *Proceedings of Smell, Taste, and Temperature Interfaces* (STT2021).

Morag, M. (2003). French Harems: Images of the Orient in Cosmetic Advertisements, 1750-1815. *Journal of the Western Society for French History*, 31.

Tonelli, S. and Menini, S. (2021). FrameNet-like Annotation of Olfactory Information in Texts, *The 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature 2021*.

Tullett, W. (2019). Smell in Eighteenth-Century England: A Social Sense. Oxford: Oxford University Press.

The Concept of Nature in German Romanticism: An Approximation

Uglanova, Inna
inna.uglanova@tu-darmstadt.de
Technical University of Darmstadt, Germany

Gius, Evelyn
evelyn.gius@tu-darmstadt.de
Technical University of Darmstadt, Germany

Introduction

Nature plays a particular role in the Romantics' worldview. This concept is characterised by a shift in emphasis from nature as a passive principle (cf. natura naturata by Schelling, 1994) to nature as a subject, an active, creative principle (natura naturans ibid.). The unifying force between these states is the spirit ('Weltseele'). This leads to the dissolving of established boundaries and oppositions between a human being and nature in romanticist texts (cf. Wanning, 2005). Romantics transform the vertical world order into a horizontal one and create a worldview where human being forms a unity with the world around them. Among other, these transformations manifest themselves in the anthropomorphisation of nature.

Our approach is therefore to find out how nature is represented in Romanticist texts by analysing whether it is depicted as an active entity and related to the concept

of human being, both in terms of anthropomorphism and the possible semantic projection as a holistic concept. While we acknowledge that nature is a complex concept with blurred boundaries, in this contribution we would like to demonstrate how one can gain insights with a relatively simple approach focused on nature as 'nature' and comparing Romanticism to other epochs.

Data

For our analysis, we compiled two corpora. Our main corpus contains 90 novels (10,511,420 tokens) from 1780 until 1850, coinciding with the epoch of German Romanticism. A comparative corpus with 102 novels (10,423,812 tokens) from Realism (56 novels) and Naturalism (46 novels) published from 1880 until 1900 was taken from the d-Prose corpus (Gius et al., 2020). This corpus was selected for a contrastive analysis since, in Naturalism and Realism, writers saw their conception of the world as the antithesis of Romanticism.

Methods

For exploring nature's agency in the sociological sense, i.e., the ability to act independently, we took the grammatical position of 'nature' as a proxy and parsed both corpora using spaCy (Honnibal et al., 2021). The subsequent analysis of the possible anthropomorphic representation of nature in Romanticism and Naturalism is based on bigram collocations of 'nature' identified with NLTK (Bird et al., 2009). The strength of association between collocates was measured by log-likelihood for collocations occurring more than three times. As an exploration of the semantic dimensions of nature, we finally constructed word embedding models for nature and human beings in Romanticism and Naturalism using word2vec in gensim (Mikolov et al. 2013; Rehurek, 2021) and visualised them with the t-SNE-algorithm (Maaten and Hinton, 2008) implemented by scikit-learn (Pedregosa et al., 2011).

Selected Results

We now sketch some key results from our project. Table 1 shows the most frequent verbs in sentences where 'nature' is used in the subject and thus, from a grammatical perspective, in an active position. In both cases, nature seems to be conceptualised as acting human-like. However, the proportion of sentences with 'nature' as subject is higher in Romanticism with 0,38% (i.e., 1,382 sentences out of 363,318) against 0,06% (305 sentences out of 545,023).

Romanticism			Realism & Naturalism		
Verb	Frequency	z-Score	Verb	Frequency	z-Score
geben give	37	9,09	scheinen seem	13	7,08
scheinen seem	37	8,30	machen make	9	4,59
sagen say	34	6,98	sagen say	8	3,96
sehen see	29	6,72	lassen let	7	3,34
wissen know	28	6,19	verlangen demand	7	3,34
lassen let	26	5,14	kommen come	6	2,72
machen make	22	4,61	wissen know	6	2,72
sprechen speek	20	4,61	liegen lie	6	2,72
haben have	20	3,55	geben give	5	2,09
stehen stay	16	3,29	haben have	5	2,09

Table 1:
Most frequent verbs in sentences with 'nature' as subject in Romanticism and Naturalism/Realism

Interestingly, we found more human-related words in bigrams in the Naturalist's subset than in Romanticism (see highlighted words in Table 2). However, the semantics of these words refer rather to human nature in the sense of character traits than to nature.

Romanticism			Naturalism		
Lemma		Log-Likelihood	Lemma		Log-Likelihood
der the	Natur nature	7.337,92	der the	Natur nature	2.316,92
menschlich human	Natur nature	677,30	**menschlich** human	Natur nature	255,64
ganz whole	Natur nature	424,95	ganz whole	Natur nature	169,99
schön beautiful	Natur nature	241,66	und and	Natur nature	159,23
Natur nature	und and	158,45	mein my	Natur nature	158,19
und and	Natur nature	136,64	angelegt enclosed	Natur nature	150,39
in in	Natur nature	133,09	von from	Natur nature	149,15
göttlich divine	Natur nature	130,12	Natur nature	selbst itself	141,71
ich I	Natur nature	115,16	**innerst** intimate	Natur nature	141,26
Natur nature	gemäß according to	108,74	**kräftig** strong	Natur nature	140,88
gütig kind	Natur nature	96,51	**leidenschaftlich** passionate	Natur nature	138,77
ein a	Natur nature	93,47	**zweiter** second	Natur nature	124,12
von from	Natur nature	90,54	**gesund** healthy	Natur nature	117,63
leblos lifeless	Natur nature	85,88	Natur nature	und and	113,76
Natur nature	selbst itself	85,26	Mutter mother	Natur nature	111,53

Table 2:
Collocations with 'nature' in Romanticism and Naturalism

The word2vec analysis (see Table 3) seems to confirm this. In the compared datasets, different concepts of 'nature' are contextualised. While the Romantics aestheticise **nature**, Naturalists refer to 'nature' as **human** characteristics and thus as **human** nature.

Romanticism		Naturalism	
Word Vector	Similarity Value	Word Vector	Similarity Value
Menschheit humanity	0,79	Selbstsucht selfishness	0,76
Wirklichkeit reality	0,78	Phantasie fancy	0,76
Schönheit beauty	0,78	Einbildungskraft imagination	0,76
Harmonie harmony	0,77	Auffassung conception	0,75
Poesie poetry	0,77	Schönheit beauty	0,75
Schöpfung creation	0,76	Eitelkeit vanity	0,75
Tugend virtue	0,76	Sinnlichkeit sensuality	0,75
Kunst art	0,74	Reinheit purity	0,75
Religion religion	0,74	Tugend virtue	0,73
Gottheit divinity	0,73	Empfindung feeling	0,73
Vernunft reason	0,72	Anmut grace	0,72
Macht power	0,71	Gesinnung attitude	0,71
Armut poverty	0,70	Herzensgüte kindheartedness	0,70
Wissenschaft science	0,70	Intelligenz intelligence	0,70
Unschuld innocence	0,70	Persönlichkeit personality	0,70

Table 3:
Similar word vectors to the keyword 'nature'

Finally, the t-SNE-visualisations with the projections of the vectors 'human being' and 'nature' into the same two-dimensional space make the different conceptions of Romanticism and Naturalism visible (cf. Fig. 1 and 2).

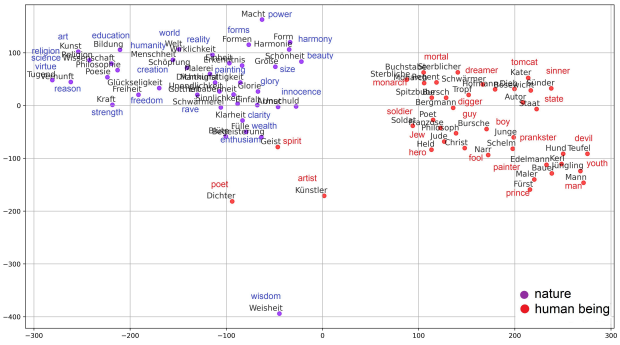


Figure 1:
The t-SNE projection of the word vectors 'nature' and 'human being' (Romanticism)

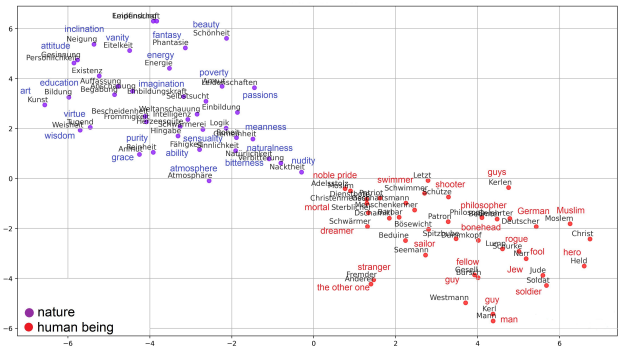


Figure 2:
The t-SNE projection of the word vectors 'nature' and 'human being' (Naturalism)

While the–more compact–cluster of the 'nature' vector in Romanticism is built by words belonging to the semantic field of human being or aesthetical, metaphysical concepts, in Naturalism, it is more directed towards human-related, often moral or character-related, properties. The 'human being' clusters, on the contrary, address humans in their social roles in both epochs. Moreover, the 'nature' cluster for Romanticism seems to point at the 'Universalpoesie' in Romanticism with its ideal of merging the fields of life, science, art and nature.

Bibliography

Bird, S., Klein, E. and Loper, E. (2009). *Natural Language Processing with Python*. Beijing; Cambridge [Mass.]: O'Reilly.

Gius, E., Guhr, S. and Adelmann, B. (2020). d-Prose 1870-1920 Zenodo doi:10.5281/zenodo.4315209. https://zenodo.org/record/4315209 (accessed 7 April 2022).

Honnibal, M., Montani, I. and Van Landeghem, S. (2021). *SpaCy · Industrial-Strength Natural Language Processing in Python*. Python and Cython Berlin: Explosion https://spacy.io/ (accessed 7 April 2022).

Maaten, L. van der and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, **9**(86): 2579–605.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, vol. 26. Curran Associates, Inc. https://papers.nips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html (accessed 7 April 2022).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011). Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research*, **12**: 2825–30.

Řehůřek, R. (2021). *Gensim: Topic Modelling for Humans*. https://radimrehurek.com/gensim/models/word2vec.html (accessed 7 April 2022).

Schelling, F. W. J. von (1994). *Ideen zu einer Philosophie der Natur: (1797)*. (Ed.) Durner, M. (Historisch-Kritische Ausgabe). Stuttgart: Frommann-Holzboog.

Wanning, B. (2005). *Die Fiktionalität der Natur: Studien zum Naturbegriff in Erzähltexten der Romantik und des Realismus*. Berlin: Weidler.

Reading *The Unknown* with a Network Mapping Device: Graph Data Visualization for Hyperfiction Works

Viehhauser, Gabriel
viehhauser@ilw.uni-stuttgart.de
University of Stuttgart

Schlesinger, Claus-Michael
claus-michael.schlesinger@ilw.uni-stuttgart.de
University of Stuttgart

Hein, Pascal
pascal.hein@ilw.uni-stuttgart.de
University of Stuttgart

Blessing, Andre
andre.blessing@ims.uni-stuttgart.de
University of Stuttgart

Ulrich, Mona
Mona.Ulrich@dla-marbach.de
Deutsches Literaturarchiv Marbach

Because of their hypertextual structure, the use of topological metaphors for the analysis of hyperfiction works appears to be obvious: Reading hyperfiction seems like walking through a maze (Schmundt, 1994). However, discussion about mapping a structure that is otherwise perceived in a processual manner, namely by (non-)linear reading, has been abundant (Ciccoricco, 2004).

For our approach to *The Unknown*, a hyperfiction novel by William Gillespie, Frank Marquardt, Scott Rettberg and Dirk Stratton (Gillespie et al:, 1998-2002), we use computational methods to extract network data from the site to visualize its hypertext structure. As web pages tend to change over time, our software Warc2graph focuses on WARC files, which are a standardized format for web archiving (IIPC, n.d.), containing data pertaining to the complete communication between a client and the server. Given a WARC file, Warc2graph extracts references from the archived object, encompassing any HTML tag that contains reference information, such as hyperlinks in an <a> tag, embedded media, automatic forwarding defined in a <meta> tag etc. We use three different methods for reference extraction: a) evaluate the WARC header, b) parse the HTML workload, c) parse the Document Object Tree of the workload utilizing Selenium, a software controlled browser engine (see Ulrich et al., 2021, for a detailed description of the methodology applied by Warc2graph).

The extracted reference information can be understood as the graph structure of an archived web object. We visualize this structure as a network showing each resource as a node and the references between resources as edges. Graph data analytics provide different measures that can be used to gain insights into the object's structure, e.g. in order to describe the functional role of certain nodes in the network. The network thus provides a structural overview of the hypertext work, which opens up new possibilities for reading hyperfiction, which, by the nature of its medium often defies the rules of linearity: Besides the sequential connection of network nodes, the graph data analysis e.g. reveals possible end points of the story line(s), which can be manifold in hypertext. It is also possible to detect central passages that can be crucial for the decisions the reader has to make, as she finds her own way through the text, since it is likely that these passages are represented by nodes that show characteristic patterns.

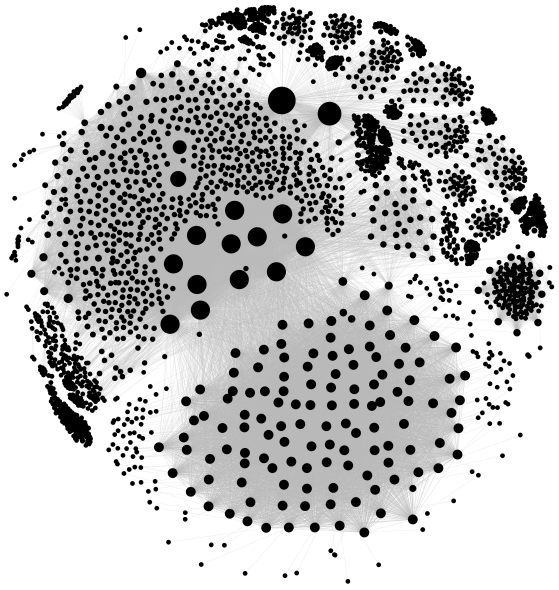


Figure 1:
Internal reference structure of The Unknown, *node size pertaining to degree*

Figure 1 shows such a network of *The Unknown*.[1] The nodes are sized according to their degree, revealing pages that appear of central importance for the site as well as clusters of interlinked documents. The network thus establishes a spatial text-view, which seems particularly apt for *The Unknown*, a text that very explicitly emphasizes its status as a non-linear hypertext (Rettberg, 2019: 80-83). By sketching out the fictional narrative of a book tour of the authors through the US the novel itself becomes a road trip through the unknown territory of hypertext, an "encyclopedic novel in a network environment" that evokes "worlds that never quite attain or even aspire to a coherent 'entirety'" (Ciccoricco, 2007: 129). Whereas these 'worlds' can be related to the clusters of the network, the graph metrics also seem to be instructive for the characterization of single nodes: E.g., one of the nodes with the highest degree refers to the page 'default.html', which contains an ironic metafictional discussion on the unsuitability of the concept of a home page for hypertext literature, which, strictly speaking, as a single starting point contradicts the principle of nonlinearity. Revealingly, the node of default.html has 129 outgoing edges, which reflects the fact that the page is ironically interspersed with links that in fact could make the site apt to serve as an explanatory starting page for the novel. However, the authors have deliberately undermined this function by 'hiding' the page in the jumble of HTML-documents, since only five ingoing links refer to the node.

The analysis also reveals that an approximation of textual structures with Warc2graph has to reflect technical constraints that could affect its results: In *The Unknown*, the main navigation was copied into each of the over 700 HTML-documents, automatically providing the navigation elements with a network degree of 759 and clustering these nodes in our graph. Another cluster with highly interlinked elements consists of 123 HTML-pages with watercolor drawings by Katie Gilligan, which accompany the text as sort of a diary. All pages are interlinked by a calendar on each page, leading to a similar degree for each document. Thus, technical features like site navigation could reveal structural similarities of nodes, but also have to be taken into account as potential biases for further interpretation.

If reading a hypertext is like walking through a maze, the graph might be near to a map of the maze. However, in order to construct a network, numerous parameters need to be set, which implies decisions that are not independent from reading and interpreting the text. Thus, rather than provide a definite map of the maze, our networks add additional views on structural qualities of The Unknown.

Bibliography

Ciccoricco, D. (2004). Network Vistas: Folding the Cognitive Map. *Image & Narrative,* 8. http://www.imageandnarrative.be/inarchive/issue08/daveciccoricco.htm#013 (accessed April 27, 2022).

Ciccoricco, D. (2008). Reading Network Fiction. Tuscaloosa: Univ. of Alabama Press.

Gillespie, W., Marquardt, F., Rettberg, S. and Stratton, D. (1998). *The Unknown*. https://unknownhypertext.com/ (accessed April 27, 2022).

IIPC. n.d. The WARC Format. https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/(accessed April 27, 2022).

Rettberg, S. (2019). Electronic Literature. Cambridge, UK ; Medford, MA: polity.

Schmundt, H. (1994). Autor Ex Machina: Electronic Hyperfictions: Utopian Poststructuralism and the Romanticism of the Computer Age. *AAA: Arbeiten Aus Anglistik Und Amerikanistik* 19 (2): 223–46. http://www.jstor.org/stable/43023678 (accessed April 27, 2022).

Ulrich, M., Schlesinger, C.-M., Blessing, A., and Hein, P. (2021). Networks of Net Literature - Modelling, Extracting and Visualizing Link-Based Networks in the DLA Corpus of Net Literature. https://elmcip.net/node/16380 (accessed April 27, 2022).

Notes

1. We provide the extracted network data together with a copy of the Warc2graph software package via Zenodo, DOI: 10.5281/zenodo.5773099

Spatial accessibility of China railway transportation network in the first half of the 20th century

Wang, Changsong
wchs@pku.edu.cn
Peking University, China, People's Republic of

Duan, Yunxin
yunxin0904@qq.com
Peking University, China, People's Republic of

Modelling Time Ontology in Ancient Chinese Texts

Wang, Linxu
wanglinxu@pku.edu.cn
Department of Information Management, Peking University; The Center for Digital Humanities at Peking University, Beijing, China

Tong, Wei
weitong@pku.edu.cn
Department of Information Management, Peking University; The Center for Digital Humanities at Peking University, Beijing, China

Wang, Jun
junwang@pku.edu.cn
Department of Information Management, Peking University; The Center for Digital Humanities at Peking University, Beijing, China

Abstract This paper aims to build an ontology of temporal concepts to describe the temporal properties of resources in the Chinese history. To represent ancient Chinese time and transform that time in history into our familiar timestamp, we present the Ancient Chinese Traditional Time Ontology (ACTO). ACTO focuses on the Calendars reference system and Clocks reference system and splits the former one into lunar year representation subsystems and lunar date representation subsystems, which makes the time can be expressed and transformed more clearly and convenience. [1]

Keyword digital humanities, ancient Chinese time, time ontology, Chinese calendar

1. Introduction

Temporal information is one of the essential components in humanity researches in the various disciplines. However, because of the unique representation approach of ancient Chinese time, it can't analyze and compute time directly. The most significant part of this approach is the traditional Chinese calendar, a type of lunisolar calendar, which is totally different from the Gregorian calendar. The traditional Chinese calendar stressed importance about the existence of the cycle about time when recording time by characters. For thousands of years, there was neither interruption nor disorder since the Chinese have used 60 symbols in the stem-branch system to mark time. But this tradition makes it is more difficult to identify exact date for repeated temporal information. For instance, A.D.2021 and A.D. 1961 can be called by one word, which is Xin-chou (辛丑 in Chinese).

It orders to translate the traditional Chinese calendar into the Gregorian calendar which is more familiar for us and transform the temporal cycle into a line, which is more convenient for analysis and computing. It's necessary and crucial to apply ontology. Ontology can provide a rich "schema" underlying data as well as the terminological and semantic basis for dramatic improvements in data application (Kendall, McGuinness 2019), which is the necessary support for organization of data. By modelling time ontology, we can use a general framework to determine, represent, sort and manage temporal information in a historical resource to improve searching, navigating, annotating, indexing and analyzing. However, there are few studies focus on the modelling and application of the traditional Chinese time.

In this paper, we address the Ancient Chinese Traditional Time Ontology (ACTO). ACTO is an ontology of temporal concepts to describe the temporal properties of resources in the Chinese history texts. Time positions and durations may be expressed using either the conventional calendar (Gregorian calendar) and clock, or traditional Chinese calendar. The reset part of the paper is organized as follows. Section 2 introduces the related work and state of the art. Section 3 describes the construction of the ACTO, including objectives and methodology. And in the Section 4, we summary our work and talk about the future work.

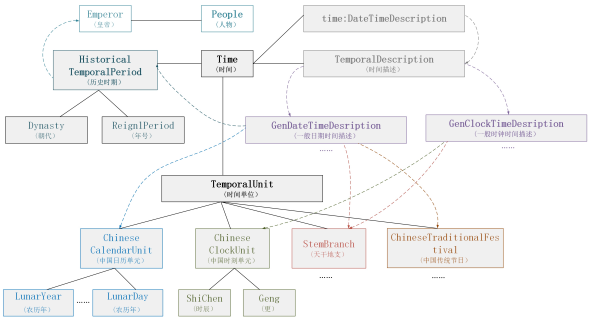
2. Related Work

Many efforts have been made on building explicit time ontologies, such as the DAML ontology of time(Daml n.d.), the time ontology in OWL(W3C n.d.), KSL time ontology(KSL n.d.), the time ontology in KIF(KIF n.d.), Times and Dates in Cyc knowledge base(CYC n.d.), the time of DAML-S(DAML n.d.), the temporal portion of IEEE Standard Upper Ontology(IEEE n.d.), and other works(Schilder, Katz & Pustejovsky 2007), (Schreiber 1994). OWL-Time is one of the most common and generalized models for expressing temporal information. It provides a lightweight model for the formalization of temporal objects, based on Allen's temporal interval calculus(Allen 1983). Since OWL-Time only allows for Gregorian dates and times, many other calendars and temporal reference systems used in the particular culture and scholarly contexts must look elsewhere. The traditional Chinese calendar is one of those calendars. Therefore, it's crucial to express the traditional temporal information encoded by literals correctly and translate it to the timestamp for calculating.

There're some studies talked the Chinese time ontology. Zhang et al. presented a Chinese time ontology involving temporal entities or temporal properties (Zhang et al. 2011), but they neither explained the structure of various concepts nor given evaluation to show the feasibility of the time scheme. Zou et al. categorized the usage of Chinese calendar under the framework of owl time (Zou et al. 2011). But the concepts in the framework are too wide, and ignores the mapping of Chinese calendar and the other calendar. Therefore, modelling Chinese time not only need to build a knowledge framework in the form of an open ontology but also need to provide a practice of that framework.

3. Objectives and Methodology

This study proposes the Ancient Chinese Traditional Time Ontology(ACTO) for modelling and publishing time entities by using OWL. By examining the distinctive features of ancient Chinese traditional time from the western time scale, the ACTO ontology has two aims. The first one is to build a knowledge representation of the time recorded in the ancient Chinese document in the form of an open ontology. The second one is to provide a practice of transformation of Chinese traditional calendar. In Figure. 1 we illustrate a part of the ontology.



A sample of the ACTO ontology

We followed a term-and-characteristic guided approach (Wei et al. 2021) to construct the ACTO ontology. The term-and-characteristic guided approach includes seven steps as following. But in the step 4, we defined time concept depended on the existed temporal reference systems rather than relying on essential characteristics.

Step 1: Identify the scope of the domain and the objectives. The aim of this step is to define the scope of the project and its objectives. In this step, we defne the scope of ACTO is the temporal information recorded in the historical document, not the Chinese traditional Calendars itself.

Step 2: Identify terms and objects. We collect records about time from history documents and select the terms and objects in those records.

Step 3: Identify the essential characteristic. One of methods to identifye essential characteristics is based on a morphological analysis of Chinese terms whose characters directly express knowledge of the denoted objects. Terms are composed of a set of characters of which the last one corresponds to the type of time and the others, called modifiers, precise the type. For example, a year can be described as the following order: Dynasty's name + Emperor's Title + Reign's Title + Ordinal Number, like '唐太宗貞觀元年'. This object belongs to the type of the description of general year time.

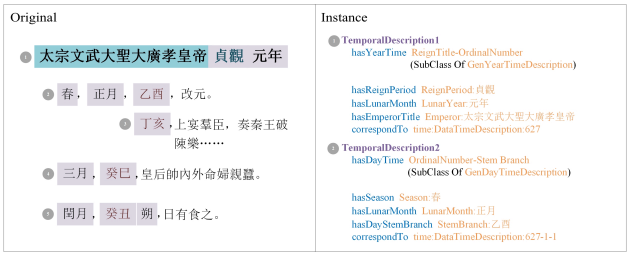
Step 4: Term-guided approach for defining concepts based on essential characteristics. Concepts are defined as meaningful combinations of essential characteristics. We apply the classification of temporal reference systems provided in ISO 19018 to describe temporal information, especially Calendars and Clocks (Cox 2016). Besides, some vocabulary from OWL-Time are used in our ontology for expressing facts about topological (ordering) relations among instants and intervals. And finally, we summary and distinguish the objects about Clocks and Calendars. Although the construction of ontology to describe temporal in the ancient China is our aim, time is not only one needed to be considered. For instance, Emperor is critical for the representation of histoical period.

Step 5: Building ontology by tools. Protégé was used for developing the ACTO, while OWL was chosen as a modeling language.

Step 6: Integration. We could consider reuse of definitions already built into other ontologies instead of starting from scratch.

Step 7: Equations.

ACTO contains 46 classes and 21 object properties. By 44 subclasses, it can describe eighteen different kinds of traditional Chinese temporal records. In Figure. 2, we use some temporal records in the *Zi Zhi Tong Jian* (the another name of this book is *History as a Mirror*) as a sample to show the usage of ACTO.



A sample of usage

4. Conclusion

Building a model of time ontology is a significant task in the field of information organization and representation. Especially with regards to an ancient time, when many aspects of historical events are highly related to other time points. It's necessary and valuable to express time and transform that time in history into our familiar timestamp.

To express and transform that, we present the Ancient Chinese Traditional Time Ontology. ACTO focuses on the Calendars reference system and Clocks reference system. And it splits the former one into lunar year representation subsystems (including nine kinds of representation methods about year) and lunar date representation subsystems (including nine kinds of representation approaches about day). Therefore, the time can be expressed and transformed more clearly and convenience. Meanwhile, we also value the independence about time entities and any calendar, culture and language by separating the calendar system and the time system. It can provide a reference for cross-cultural time translation.

Still, our work has many limitations which we will address in the future. In the next step, our work will carry out in three different directions, including enriching individuals; visualizing the ontology for query; and address a generalized proposal for evaluating the time ontology.

Bibliography

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. Communications of the ACM, 26(11), pp.832-843.

Cox, S. J. (2016). Time ontology extended for non-Gregorian calendar applications. Semantic Web, 7(2), pp.201-209.

Daml-time homepage. Available from: http://www.cs.rochester.edu/~ferguson/daml. [10 December 2021].

Daml-s. Available from: http://www.daml.org/services/daml-s/0.9. [10 December 2021].

Time and dates. Available from: http://www.cyc.com/cycdoc/vocab/time-vocab.html. [10 December 2021].

Time ontology needed. Available from: http://suo.ieee.org/SUO/SUMO/index.html. [10 December 2021].

Time ontology in owl. Available from: http://www.w3.org/TR/owl-time. [10 December 2021].

Time ontology in kif. Available from: http://www-cs-students.stanford.edu/~omart/timeont.html. [10 December 2021].

Ksl-time. Available from: http://www.ksl.stanford.edu/ontologies/time. [10 December 2021].

Schilder, F., Katz, G., & Pustejovsky, J. (2007). Annotating, extracting and reasoning about time and events. In Annotating, extracting and reasoning about time and events. Springer, Berlin, Heidelberg, pp. 1-6.

Schreiber, F. A. (1994). Is time a real time? An overview of time ontology in informatics. Real time computing, pp. 283-307.

Kendall, E. F., & McGuinness, D. L. (2019). Ontology engineering. Synthesis Lectures on The Semantic Web: Theory and Technology, 9(1), pp. i-102.

Wei, T., Roche, C., Papadopoulou, M., & Jia, Y. (2021). Using ISO and Semantic Web standard for building a multilingual terminology e-Dictionary: A use case of Chinese ceramic vases. Journal of Information Science, 01655515211022185.

Zhang, C., Cao, C., Sui, Y., & Wu, X. (2011). A Chinese time ontology for the Semantic Web. Knowledge-Based Systems, 24(7), pp. 1057-1074.

Zou, Q., & Park, E. G. (2011). Modelling ancient Chinese time ontology. Journal of Information Science, 37(3), pp. 332-341.

Notes

1. This research is supported by the NSFC project "the Construction of the Knowledge Graph for the History of Chinese Confucianism" (Grant No. 72010107003).

Quantitative understanding of the evolution of Seo Jeong-Ju's poetic world: Keywords, topics, and sentiments.

Wang, Sungpil
wangsp0317@kaist.ac.kr
KAIST, Korea, Republic of (South Korea)

Park, Juyong
juyongp@kaist.ac.kr
KAIST, Korea, Republic of (South Korea)

In this paper we provide a quantitative, computational analysis of the evolution of the poems of Seo Jeong-Ju (1915–2000), Korea's eminent 20th-century. The publications of his books of poems are traditionally understood to have marked the evolution of his poetry, which we explore in this work using modern data-processing techniques for capturing his changing compositional styles. The corpus consists of 98290 words extracted from a total of 950 poems that naturally vary in length. The number of words in a poem ranges from 9 to 699 with average and standard deviation 109.2 ± 94.2.

We start by computing the changes in the frequency of vocabulary used. We focus particularly on the words related to colour and the names of Korea's historical kingdoms conjectured as forming the identity of Seo's poetry in past qualitative studies. We count the frequencies of twelve colours and eight historical Kingdoms and see how they change over time that point to the trends in Seo's compositional style. (1) Colour: There are six colors with the highest dominance, all of which belong to the Oche(五彩) of the East. The dominance of blue and green in <The essence of Silla> (Seo, 1961) has a high point, and the dominance of red in <Jilmajae myth> (Seo, 1975) has a low point. Over time, Seo Jeong-Ju's poetic world bleaches and the influence of black and white increases. (2) Kingdom: Throughout the entire period, he mainly uses words related to Silla and Joseon. In the early days, his attention is focused on Silla, and Joseon's dominance increases from <The essence of Silla> (Seo, 1961) to <Jilmajae myth> (Seo, 1975). The results of the analysis of the keywords express support or add comments to the results of qualitative studies dealing with Colour and Kingdom.

We also analyse Seo Jeong-Ju's poetry using the tools of computational linguistics, namely Topic Modeling and Sentiment Analysis. Topic Modeling refers to a set of methods (including Structural Topic Modeling (Margaret et al., 2019) we use here) that discover groups of important words appearing frequently together in a text and allow us to identify its topic. We use diagnostic values to determine the number of topics as eight and extract each topic in the form of a noun list. Each topic is independent of each other, and the nouns that make up them belong mainly to the broad classification of family, country, women, and nature. In addition, STM provides whether the subjects are strengthened or weakened as a function of time. Through this, we seize that the four topics are strengthened and the other four topics are weakened, especially The third of the eight extracted themes appeared in <The essence of Silla> (Seo, 1961) and continued to be strengthened, and eventually became the most important theme in the literary world of Seo Jeong-Ju. Furthermore, We cross-reference the topics with other poets' and critics' writings from the era to understand what influenced the topics. Our findings support qualitative research results (Jeong, 2005; Nam, 2011; Yun, 2013) on the timing of the emergence of "Silla spirit", a core subject matter of Seo Jeong-Ju, and the persistence of this topic.

Sentiment Analysis refers to determining the positivity or negativity of the sentiment of a given text. It is well accepted that an artist's mental state and their creations are intimately related (Terry, 2008); the emotional air of a work is naturally aroused by the artist's creative choices, intentional or unintentional, reflecting their thoughts or feeling of the moment. The evolution in the sentiment of poems, therefore, could also indicate a meaningful characteristic of the works. In order to comprehend the changes in emotions that make up the poetic world of Seo Jeong-Ju, We search for emotional words in each collection of poems and calculate the emotional score of each collection of poems using the *KNU Korean emotional dictionary* (Park et al., 2018) that illustrates multiple polarities (-2, -1, 0, +1, +2) as lexicon. We discover that since <Jilmajae myth> (Seo, 1975), the sentiment of his literary creation has been overturned from negativity to positivity. These findings support specific qualitative research results (An, 2011; Heo, 2009) that post - <Jilmajae myth> (Seo, 1975) consist of small allegories or stories affirming the subject consciousness of previous poems, and that these small adaptations contain feelings of resignation and compliance.

Due to their mostly qualitative nature, existing individual classical studies on Seo Jeong-Ju have focusing on only a partial, limited set of materials and periods. Our work showcases the ability of modern computational tools to take advantage of the increasing availability of large-scale digitized data to verify the findings from traditional studies, and propose novel and interesting questions.

Bibliography

An, H.-S. (2011), 'The study of the latest poetry of jung joo soe'.

Heo, Y.-H. (2009), 'A study of seo cheong-ju's poetry - on focus to latest period'.

Jeong, H.-G. (2005), 'The imagination and ideology of seo jeong-ju's latter poems', Journal of academic conferences of International Association of Language Literature 2005(2), 126–131.

Nam, K.-H. (2011), 'Reviewing seo jeongju's poetics of shilla-spirit (its ethical consciousness and political unconsciousness)', Korean Culture 54.

Park, S.-M., Na, C.-W., Choi, M.-S., Lee, D.-H. and On, B.-W. (2018), 'Knu korean sentiment lexicon: Bi-lstm-based method for building a korean sentiment lexicon', Journal of Intelligence and Information Systems 24(4), 219–240.

Roberts, M. E., Stewart, B. M. and Tingley, D. (2019), 'Stm: An r package for structural topic models', Journal of Statistical Software 91, 1–40.

Rustin, T. A. (2008), 'Using artwork to understand the experience of mental illness: Mainstream artists and outsider artists', GMS Psycho-Social Medicine 5.

Seo, J.-J. (1961), *The essence of Silla,* Jeongeumsa.

Seo, J.-J. (1975), *Jilmajae myth*, Iljisa.

Yun, J.-W. (2013), 'The aesthetics of action in jilmajaeshinhwa', Journal of Dong-ak Language and Literature 61, 367–396.

Affective Writing in Ming Dynasty *Huaben* Stories — A Topic Modeling Study

Wang, Yiwen
11804048@zju.edu.cn
Zhejiang University, People's Republic of China

Kurzynski, Maciej
makurz@stanford.edu
Stanford University, United States of America

Topic modeling is a text mining technique used to discover latent semantic dimensions in a corpus of texts. In our paper, we employ the LDA topic modeling algorithm to gain new insights into the language of *huaben* narratives (short- or medium-length stories or novellas in vernacular language) written in early modern China during the Ming and Qing Dynasties. Our corpus consists of 633 stories assembled from multiple digital databases according to Ouyang Daifa's 欧阳代发 *History of Huaben Literature* 《话本小说史》. Our quantitative analysis discovers a number of computational topics related to the discourse of *qing* (情, often translated into English as "Feeling," "sentiment," or "sensibility"). We combine topic modeling (distant reading) with close reading in order to trace the evolution of narrative patterns and examine the usage of vocabulary during the so-called "cult of *qing*" in the late Ming Dynasty. We show how the discourse of *qing* became increasingly specialized and sophisticated, and how this development mirrored the evolution of sensibilities of both authors and readers of the *huaben* stories. Specifically, we explore the ways in which writers combined the themes of love, erotic desire, and marriage in their works. The ornate vocabulary of poetic love was used to "set the tone" through opening poems, to structure the story's plot via extra-diegetic transitions, and to provide a narrative frame to an otherwise unsophisticated pornographic content. The erotic vocabulary, on the other hand, which developed as a vernacular counterpart to the more abstract poetic discourse, was frequently employed to describe bodily experiences but also the "negative" or "evil" characters. The marriage vocabulary, finally, was largely unrelated to the poetic discourse. We argue that the presence or absence of erotic vocabulary served as an implicit moral judgement passed on literary characters by the narrators. More generally, our project shows the ways in which computational criticism can enrich traditional literary scholarship by drawing our attention to the very materiality of texts and the words that constitute it.

Curuinsi Project: A lexical database for preserving Tikuna language

Wicht, Bertil
bertil.wicht@unil.ch
RHYZOM SA, Switzerland

Picca, Davide
davide.picca@unil.ch
Université de Lausanne, Switzerland

Introduction

In the field of Natural Language Processing, the concentration of technological efforts focuses on widely spoken languages. In this regard, the vast majority of spoken languages can be characterized as under-resourced

languages (Thai et al., 2019).This lack of digital resources and projects is particularly challenging for indigenous languages that are facing the threat of extinction. In order to address such a compelling issue, we spent one month doing fieldwork with native speakers, thanks to which some shortcomings emerged. In particular, the lack of technological support to help the preservation of this language.

Context

Tikuna is spoken by a relatively small group of people of roughly 20,000 people living in the Amazon. The language is slowly dying, replaced by Spanish. Despite the continuous efforts of linguists and anthropologists whose studies keep the language alive and the interest in maintaining it, at present there are no persistent and ongoing digital projects that disseminate and preserve this linguistic culture.

Despite such an observation, this work was possible thanks to some very rare resources found on the web which are:

1. a word list compiled from documents translated in Tikuna (Scanell, 2007)
2. a hunter-gatherer word database containing 647 entries [1]
3. a PDF bilingual dictionary in Spanish-Tikuna (Anderson and Anderson, 2017)

Similar to our colleagues in the Kamusi project (Martin and Radtezky, 2014), creating digital resources from digitized paper documents presents methodological challenges that cannot be neglected. As in our case, the quality of digital capture is often not the best. Therefore, for this project, we focused on the best quality document, the bilingual dictionary, to create a fully digitized resource such as an interactive database.

Methodology

The overall pipeline as depicted in Illustration 1 for the project consists in 4 main steps :

1. conversion of the PDF dictionary in an editable text format
2. correction and restructuring of text
3. annotation and extension of the database
4. assessment and analysis of the data

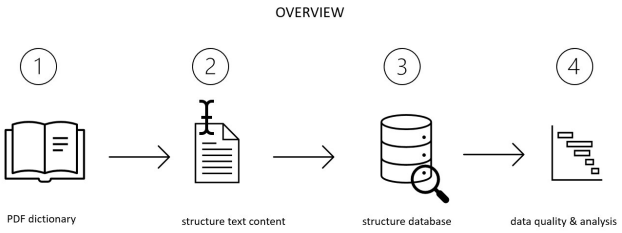


Illustration 1:
The 4 sequential steps of the text processing

For the conversion of the PDF file we employed the Adobe Acrobat OCR converter [2]. Due to the weak PDF structure and the complexity of the diacritics, the OCR output required substantial manual cleanup work during phase 2 to ensure automatic segmentation. The text file contained 5828 lines, each providing synsets for a Tikuna word or expression. Illustration 2 shows the way the original scanned document content was segmented in order to capture each distinct element. Examples were not captured due to limitations concerning regexes and lookarounds in this particular context with complex diacritics. Element referring to another Tikuna word were not kept, as the semantic relation between the terms were unclear. For those synsets lacking POS tags, we used the spacy tagger [3] to annotate them. Instead those synsets for which the dictionary provided the POS, we just standardized the tag so that it was uniform with the tagset provided by spacy. Finally, in step 4, we ran some analysis to validate the data structure and provide insights on the overall performance of the procedure.

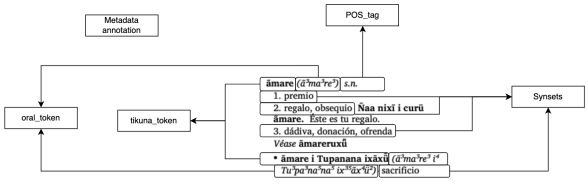


Illustration 2:
Vizualisation of the text segmentation and labelling

After the above mentioned segmentation process, we classified the synsets in four categories based on the number of synsets provided per Tikuna tokens and the number of POS tags per synsets as shown in Table 1.

Monosemic with POS tag	4094

Monosemic without POS tag	1258
Polysemic with single POS tag	544
Polysemic with multiple POS tag	42

Table 1: Overview of classified entries from the initial document

Lastly, as shown in Table 2, we assessed the data loss during the process and the missing data from the database. The only missing data regards translations and encompass all entries that did not have a word for translation but rather an explanation of the word.

Tikuna	0
Prononciation	0
Gram. Type	0
Translation	26

Table 2: Data quality assessment

The database is freely exploitable through an interactive web application, user-friendly and ready-to-use [4].

Conclusion

In a future phase of the project, we intend to extend the database and not only to transcribed entries but each entry could have all oral variants incorporated, since Tikuna is a primarily oral language and therefore many words have multiple phonetic forms. Also, since the initial document was created to facilitate indigenous reading of the Bible, most of the content is geared toward this use. Therefore, the actual content of the corpus is not representative of the language or culture of the Tikuna speakers. However, some complex cultural meaning as explained by Santos (Santos, 2013) are lacking. Better integration of indigenous ontologies is needed to ensure inclusivity and representativeness. By doing so, we can build cultural data models with other ontologies that are more inclusive and representative of cultures. For this task and for the next steps of the project, collaboration with native speakers would be necessary, which we are already working for.

Bibliography

Martin Benjamin and Paula Radetzky.(2014). *Multilingual Lexicography with a Focus on Less-Resourced Languages: Data Mining, Expert Input, Crowd-sourcing, and Gamification*. en. In: 9th edition of the Language Resources and Evaluation Conference.

Anderson Doris and Lambert Anderson (dir.).(2017). *Diccionario ticuna–castellano*, Serie Lingüística Peruana. Summer Institute of Linguistics, Lima.

Abel Antonio Santos Angarita.(2013). *Percepción tikuna de Naane y Naüne: territorio y cuerpo.* Universidad Nacional de Colombia - Sede Amazonas. PhD thesis .

Kevin Scannell.(2017). *The Crúbadán Project: Corpus building for under-resourced languages*. en. In: Cahier du Central , p. 10.

Notes

1. The Hunter – Gatherer database can be found at : https://huntergatherer.la.utexas.edu/languages/language/110
2. The Adobe OCR can be obtained at : https://www.adobe.com/acrobat/how-to/ocr-software-convert-pdf-to-text.html
3. spacy is freely available at : hhtps://spacy.io
4. The application is available upon request contacting the first author

Learn-STATIC: Building Fundamental Digital Skills in the Humanities Classroom

Wikle, Olivia M.
omwikle@uidaho.edu
University of Idaho, United States of America

Williamson, Evan Peter
ewilliamson@uidaho.edu
University of Idaho, United States of America

Becker, Devin
dbecker@uidaho.edu
University of Idaho, United States of America

Thornhill, Kate
kmthorn@uoregon.edu
University of Oregon, United States of America

Hayden, Gabriele
ghayden@uoregon.edu
University of Oregon, United States of America

Static web technologies offer an exciting opportunity for DH instructors to incorporate transferable digital and data literacy skills into their classrooms, while producing low-cost, low-maintenance interactive web projects that are sustainable even for institutions with limited resources (Wikle et al., 2020). As an alternative to dynamic web applications, static websites are composed of flat HTML, CSS, and JavaScript files that are produced using a static site generator such as Jekyll, and hosted on a simple—often free—server such as GitHub Pages. The code and data that Jekyll uses to generate a site's files is typically contained in a repository that can be stored and collaboratively edited on GitHub, GitLab, or other code hosting sites.

By virtue of their being static, the site's files will remain secure and functional over the long term with little maintenance, making statically-generated sites an ideal solution for classroom projects which usually undergo a period of intense semester-long development followed by long periods of minimal maintenance. Along with these economic and time-saving benefits, teaching with static web tools allows instructors the flexibility to introduce increasingly complex concepts of data management, web development, and computing to students. Depending on how a project is framed and taught, students may encounter anything from basic tutorials on spreadsheet formulas, to Git, GitHub, Markdown, YAML, HTML, CSS, Liquid, and Jekyll. The incorporation of these technical concepts addresses a gap in DH instruction, which all too often succeeds at teaching the "buttonology" of a specific DH platform without also investing in teaching transferable digital literacy skills that empower students to think more critically about the digital projects they use (Russell and Hensley, 2017). By contrast, when students engage with static web technologies and data structures alongside the more traditional DH methods of determining and curating humanities data (Posner, 2015), they gain the invaluable experience of exploring how their data is consumed, transformed, and output in the form of substantial, interactive web projects, and learn to bring the same spirit of critical inquiry that they focus on humanities content to their understanding of the tools and processes they use to manipulate and share digital content.

Static web technology is not new to DH: static web templates such as Ed (for publishing digital editions), Wax (for creating digital exhibits), CollectionBuilder (for building digital collections), and Oral History as Data (for curating and visualizing oral histories) embody a minimal computing ethos by reducing unnecessary reliance on databases or excessive processing power to more directly meet the unique needs of a project (Gil, 2015). However, few models or example lesson plans exist that demonstrate how to incorporate static web tools and development practices into the DH classroom.

The Learn-STATIC initiative, funded by a 2021 National Endowment for the Humanities (NEH) Digital Humanities Advancement Grant and piloted by librarians at the University of Idaho and University of Oregon, aims to address this lack of resources by creating a series of open-source learning sequences for using static web tools in the DH classroom, each complete with reusable code stored in a GitHub repository, an example lesson plan, and documentation tailored specifically for instructors' and students' use in the classroom. Each learning sequence focuses on one of the following topics: digital oral histories, text analysis, digital collections, and digital project recovery, and provides a practical way forward for instructors new to using or teaching static web tools, as the step-by-step documentation for each project walks students through a variety of activities including curating spreadsheets of data, uploading data to the project's GitHub repository, and configuring and hosting the project site on GitHub Pages.

This presentation will introduce the Learn-STATIC learning sequences, report on their effectiveness in the classroom, and demonstrate how they accomplish the advantages of static web laid out above. Ultimately, we hope that Learn-STATIC will reach beyond the initial learning sequences presented here, to serve as a collaborative initiative that invites DH practitioners to share their own creative adaptations of static web tools for DH pedagogy.

Bibliography

Gil, A. (2015). The User, the Learner and the Machines We Make. *Minimal Computing*. https://go-dh.github.io/mincomp/thoughts/2015/05/21/user-vs-learner/.

Posner, M. (2015). Humanities Data: A Necessary Contradiction. *Miriam Posner's Blog*. https://miriamposner.com/blog/humanities-data-a-necessary-contradiction/.

Russell, J. E. and Hensley, M. K. (2017). Beyond Buttonology: Digital Humanities, Digital Pedagogy, and the ACRL Framework. *College & Research Libraries News*. https://crln.acrl.org/index.php/crlnews/article/view/16833.

Wikle, O., Williamson, E. and Becker, B. (2020). What is Static Web and What's it Doing in the Digital Humanities Classroom? *dh+lib*. https://acrl.ala.org/dh/2020/06/22/what-is-static-web-and-whats-it-doing-in-the-digital-humanities-classroom/.

Teaching Basic Buddhism Using a Chatbot: Evaluation and Comparison

Wong, Kwong-Cheong

kwongcheong.wong@cpce-polyu.edu.hk
College of Professional and Continuing Education, The Hong Kong Polytechnic University

Chan, Andrew Marcus

am.chan@cpce-polyu.edu.hk
College of Professional and Continuing Education, The Hong Kong Polytechnic University

Law, Shun-Man

shunman.law@cpce-polyu.edu.hk
College of Professional and Continuing Education, The Hong Kong Polytechnic University

Tse, Devi

devitcf@gmail.com
College of Professional and Continuing Education, The Hong Kong Polytechnic University

The world is becoming more and more technological thanks to the emergence of advanced technologies like Artificial Intelligence (AI). Unfortunately, technological advances themselves have not been able to bring about the solving of the fundamental problems of human existence arising from birth, aging, disease and death . What we need, instead, is penetrating wisdom. We believe that the ancient Buddhist wisdom, which is aimed at alleviating suffering and gaining happiness and ultimately at enlightenment, is still an effective antidote to the fundamental human existential problems in our time (Wright , 2017). However, learning Buddhism is not easy, for a number of reasons. One is that it contains a lot of terminology whose meaning deviates subtly from its common usage (e.g., the words "dukkha" and "suññatā" are usually, but not entirely satisfactorily, translated as "suffering" and "emptiness", respectively), rendering Buddhism hard to grasp, especially for beginners. Although technologies cannot, in the parlance of Buddhism, be the moon itself, they can act as a finger pointing at the moon. Accordingly, the goal of this project is to leverage recent AI technology to develop a chatbot to teach the basic Buddhist doctrines revolving around the Four Noble Truths (Suffering, The Arising of Suffering, The Cessation of Suffering, The Path Leading to the Cessation of Suffering). The inspiration of this idea comes from the exhilarating story in which the Buddha, upon attaining Buddhahood, returned to teach the Four Noble Truths to the Five Bhiksus, thereby enabling the latter to attain nibbāna (enlightenment).

Technically, our chatbot is a task-oriented one. It works in the narrow domain of explaining the Four Noble Truths. It is also text-based. During its conversation with the user, the chatbot asks the user concrete questions and the user answers. Based on these answers and their sentiment (positive or negative), the chatbot would recommend the user to learn a specific Noble Truth. Throughout, emphasis is put on explaining the terminology of basic Buddhism. To be more specific, the Buddhist knowledge contained in our chatbot comes from *What the Buddha Taught*by W. Rāhula (1974). This highly acclaimed book has been praised for its clarity, accessibility and reliability, and as such has been widely adopted as an introduction to the basic Buddhist doctrines.

We took a mixed method approach to evaluating our chatbot. Qualitatively, we conducted in-depth interviews with the users, soliciting their perceptions towards, and subjective experiences of, using the chatbot to learn the basic doctrines of Buddhism. Quantitatively, we extended the original Technology Acceptance Model (TAM) (Davis et al., 1989) with some external variables (e.g., Perceived Enjoyment) in order to evaluate the users' acceptance of using the chatbot to learn basic Buddhism. Based on this extended model, we designed a survey questionnaire. The focus of this paper is on reporting the quantitative findings of this survey questionnaire and their analysis – the qualitative counterparts were reported in (Chan et al., 2021).

There exists a diversity of religions, especially in Asia, that can offer wisdom to solve the fundamental human existential problems (see, e.g., Koller, 2018) . Buddhism is just one of them; there are many others, e.g., Confucianism. In the final part of this paper, we first compare our work with Cheok and Zhang (2019), which builds a chatbot to teach the basic doctrines of Confucianism, and then compare our particular approach with that of Pataranutaporn et al. (2019), which, like us, builds a chatbot to teach basic Buddhism. Joining Tan (2020) and BBC (2021), we conclude the paper with a critical discussion on exploiting AI chatbot/robot technology in religious/spiritual education.

Bibliography

British Broadcasting Corporation (BBC) (2021). TV documentary "God and Robots: Will AI Transform Religion?" at https://www.bbc.com/news/av/technology-58983047 .

Chan, A. M., Law, S., Wong, K. & Tse, D. (2021). Digital Humanities in Teaching: The Case of Bodhi Chat in Action. *The 12 th International Conference of Digital Archives and Digital Humanities* (DADH2021).

Cheok, A. D., & Zhang, E. Y. (2019). A Virtual Confucius Chatbot. In *Human–Robot Intimate Relationships* (pp. 123-151). Springer, Cham.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User Acceptance of Computer Technology: A Comparison

of Two Theoretical Models. *Management Science*, *35* (8), 982-1003.

Koller, J. M. (2018). *Asian Philosophies* (7 th edition). Routledge.

Pataranutaporn, P., Ngamarunchot, B., Chaovavanich, K., Chatwiriyachai, S., Ngamkajornwiwat, P., Ninyawee, N., & Surareungchai, W. (2019). Buddha Bot: The Exploration of Embodied Spiritual Machine in Chatbot. In *Proceedings of the Future Technologies Conference* (pp. 589-595). Springer, Cham.

Rāhula, W. (1974). *What the Buddha Taught* (Revised edition). Grove Press.

Tan, C. (2020). Digital Confucius? Exploring the Implications of Artificial Intelligence in Spiritual Education. *Connection Science*, *32* (3), 280-291.

Wright, R. (2017). *Why Buddhism is True: The Science and Philosophy of Meditation and Enlightenment*. Simon and Schuster.

From shape to culture: a computational method to extract and study the shape of vases

Yang, Yuchen
yuchen.yang@epfl.ch
Swiss Federal Institute of Technology Lausanne, Switzerland

Han, Zhitong
zhitonghan7@gmail.com
Sichuan University, People's Republic of China

Introduction

In recent years, the idea that cultural change can be described as a process comparable to organic evolution has become increasingly popular (Mesoudi, 2001). Like organic diversity, cultural diversity results from a modification-with-descent process, and some methods used to study organic diversity can be used to explore the culture. For example, the genealogical relationships among cultural artefacts (languages, folk tales, Iranian rugs) can be reconstructed using phylogenetic methods (Tehrani and Collard, 2009; Tehrani et al., 2010) and to study the evolution and ecology of cultures.

Shape, among other quantitative features, has shown great potential for such studies. Multiple works use shape to study cultural grammar and cultural diversity (Liu and Ren, 2021; Di Angelo et al., 2018). However, these attempts either are qualitative or focus on potteries in the ancient Greek time and fail to capture the pan-Asian porcelain, which is arguably the most important artefacts that connect Asian countries and the rest of the world.

This study proposed 1) a quantitative method to study how MeiPing *(a typical Chinese* porcelain type *)* has evolved in design through shape. The null hypothesis is that the more balanced shape of MeiPing would have been adopted more often in the course of cultural evolution; accordingly, the alternative hypothesis is that MeiPing would have become less balanced, turned into a delicate ornament rather than a practical vessel. 2) a workflow to automatically extract contours to bridge the gap between methods and data for such studies in the future.

Methods

Shape analysis

Since we are trying to capture the essence of MeiPing contour, but not the exact contour - that might be weird and complex due to the decorations on the vase - we performed manual landmarking on a total of 230 MeiPing to obtain their contours. Superimposition and normalization are performed using the momocs 1.3.0 (Bon homme et al, 2014) in R.

$$\text{Balance coefficient} = \min [L_1 (1 - \sin \partial), L_2 (1 - \sin ß)]$$

Formula 1.
Balance coefficient definition

To understand how practical a vase is, we introduced two indicators. The first is the shape-dependent Balance coefficient (BC) *using Formula 1*. Each parameter used for the calculation is defined and described in Fig. 1. The second is Degree of symmetry (DC), calculated using momocs.

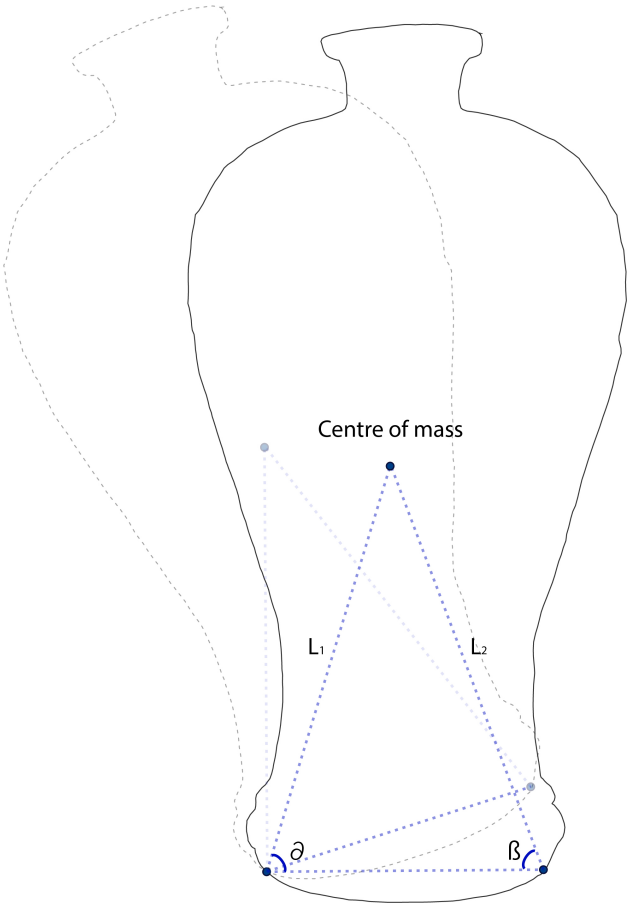


Figure 1.
Parameters definition and explanation for Balance coefficient

Contour extraction

The traditional way of extracting contours is based on binarised images, which works fine for clean and consistent inputs. In reality, artefacts images are from multiple sources, and their standards and conditions vary. The constantly changing background, contrast, and brightness make the task difficult and require constant changes to the parameters for binarisation to work(Fig. 2).

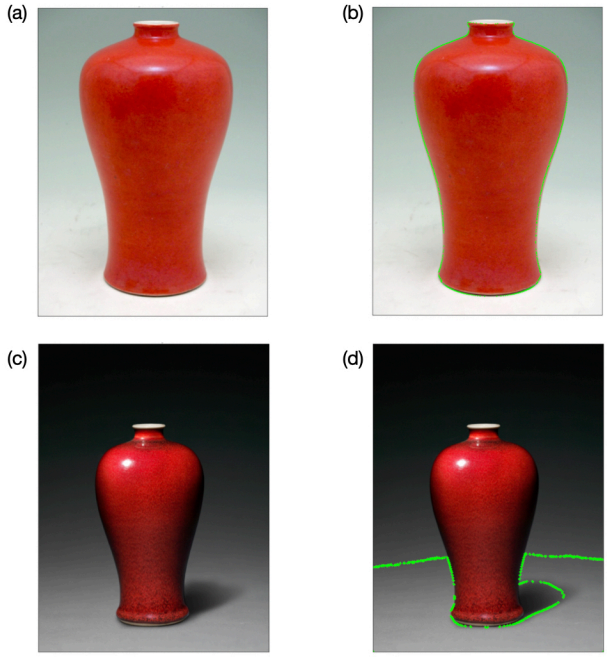


Figure 2.
(a) a good image for binarisation extraction, (b) a contour in green extracted in OpenCV, (c) a bad image for binarisation extraction, (d) a failed contour in green extracted in OpenCV

There is a rising attempt to use learning models for image segmentation that can detect and segment the objects of interest in given images. The perk of this is that it does not restrict image conditions, the segmentation can be done in much looser terms compared to binarisation. However, the go-to model Mask-RCNN (He et al., 2017) is very good at getting a rough contour but not with precision.

In this research, we integrate Pointrend (Kirillov et al., 2020), an improved model that treats segmentation as a rendering task for more precise contours, and compare it with the popular Mask-RCNN. The comparison is conducted using the strict metric Average Precision at IoU = 0.75 (AP75) in COCO detection evaluation.

Data

We have collected 230 images of MeiPing from museum digital archives to imitate real-life situations - lack consistent sources and standards, and the quality varies. To perform the quantitative analysis and validate the automated contour extraction workflow, all images are manually annotated. A synthetic dataset of 1000 vases was created for the finetuning of the machine learning model.

Results

Shape analysis results on Meiping

The results of linear regressions of both parameters, BC and DC, with the production year have a P-value < 0.001. We can then conclude that MeiPing becomes more balance and symmetrical over time (Fig. 3), therefore the null hypothesis was accepted – – the shape of MeiPing evolves to become more and more practical.

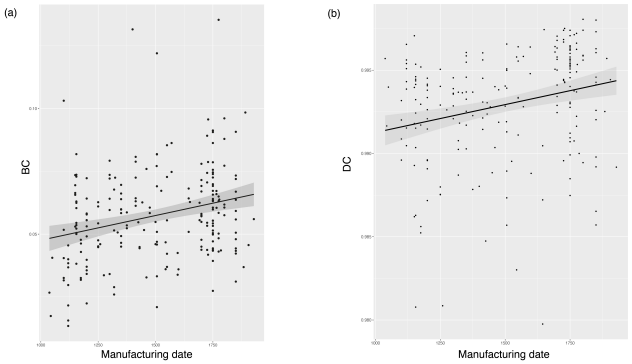


Figure 3.
(a) results of linear regression between manufacturing date and BC, (b) results of linear regression between manufacturing date and DC

Contour extraction

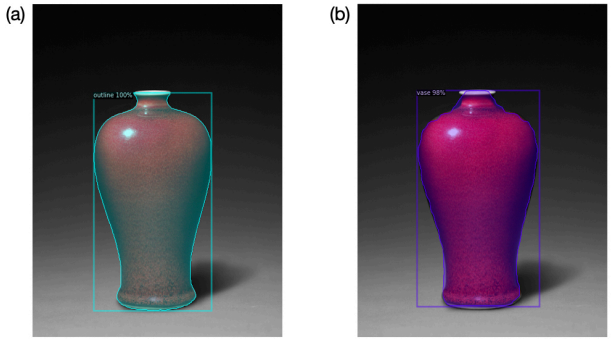


Figure 4.
(a) contour inferred by Pointrend, (b) contour inferred by Mask-RCNN

1000 synthetic images of vases against backgrounds were used to fine-tune the Mask-RCNN and Pointrend model. The inferences and evaluation were run on all manually annotated images . The Mask-RCNN achieved an AP75 of 91.10, while the Pointrend achieved an AP75 of 99.2. The differences in the AP75 value indicated the Pointrend model is better at getting the detailed mask of the object in question, hence can retrieve better contours. In future examinations of the result by eye, the Pointrend model predicted the mask area more tightly wrapping the vase, while the contour predicted by Mask-RCNN is less so (Fig. 4). Fig. 5 shows that the machine learning extracted contour reached comparable conclusions compared to the ones using manually annotated data .

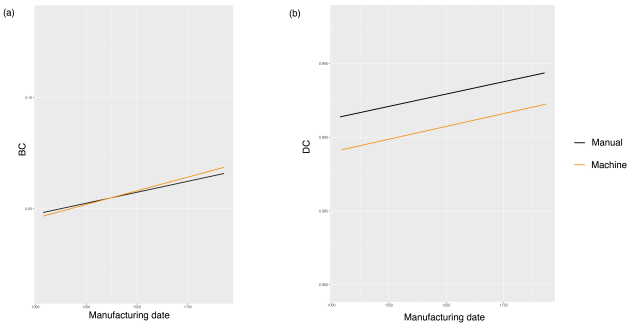


Figure 5.
comparison of results of manually annotated shape data and machine extracted contours, (a)linear regression between manufacturing date and BC, (b) linear regression between manufacturing date and DC

Conclusion

This study proposed a practical and viable quantitative method to study the functional change of vases through shape, as well as a verified workflow to automatically extract contours out of images of vases regardless of image conditions to facilitate future studies on vase shapes. The result, combined with proper metadata, allows researchers to validate existing or create new theories of cultural evolution.

Bibliography

Mesoudi, A. (2021). Cultural evolution. In Cultural Evolution. University of Chicago Press.

Tehrani, J. J., & Collard, M. (2009). On the relationship between interindividual cultural transmission and population-level cultural diversity: a case study of weaving in Iranian tribal populations. Evolution and Human Behavior, 30(4), 286-300.

Tehrani, J. J., Collard, M., & Shennan, S. J. (2010). The cophylogeny of populations and cultures: reconstructing the evolution of Iranian tribal craft traditions using trees and jungles. Philosophical Transactions of the Royal Society B: Biological Sciences, 365(1559), 3865-3874.

Liu, F., & Ren, Y. (2021). Study on Shape Design of Shouzhou Kiln Porcelain Based on Shape Grammar and

Genetic Constraints. Journal of Landscape Research, 13(4), 111-117.

Di Angelo, L., Di Stefano, P., & Pane, C. (2018). An automatic method for pottery fragments analysis. Measurement, 128, 138-148.

Bonhomme, V. et al. (2014). "Momocs: Outline Analysis Using R". In: Journal of Statistical Software 56.13, pp. 1–24.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

Kirillov, A., Wu, Y., He, K., & Girshick, R. (2020). Pointrend: Image segmentation as rendering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9799-9808).

Regularization of kinship relations to enrich the family network analysis: Case study on China Biographical Database

Yuan, Yiguo
lexcliff1023@gmail.com
Nanjing Normal Universtiy, China, People's Republic of

Li, Bin
libin.njnu@gmail.com
Nanjing Normal Universtiy, China, People's Republic of

Lu, Xuehui
1164800955@qq.com
Nanjing Normal Universtiy, China, People's Republic of

Feng, Minxuan
fennel_2006@163.com
Nanjing Normal Universtiy, China, People's Republic of

Kinship is an important issue in history studies. The kinship database is the key resource to analyze the structure and succession/evolution of families (Shang and Huang, 2018). However, the types of kinship within databases rely on the original description of kinships in the raw text. Natural languages have many kinship words to name different types of relationships. Thus, the relations extracted from raw texts cannot be directly used to definitively build family networks. As in the well-known China Biographical Database (CBDB), which contains 484,416 kinship instances, there are more than 400 types of kinship relations. In this paper, we put forward a novel method to regularize kinship relations by three basic relations: father-offspring, mother-offspring and husband-wife. All types of relations are regularized as these three basic relations to construct family networks more conveniently. Persons' information is helpful to the regularization of kinship relations and verification of kinship instances. We retain four kinds of information for each person in CBDB: ID, name, gender and year of death.

There are three steps in the regularization of kinship relations. Firstly, we extracted kinship instances via 64 kinship word types recording the three basic kinships directly and regularize them to remove the redundancy. This is because one kinship instance could come from multiple instances in CBDB. In this step, we got 121,829 basic kinship instances, including 76,597 father-offspring instances, 18,839 mother-offerspring instances and 26,393 husband-wife instances. Secondly, there are some non-basic kinship instances by which to infer new basic kinship instances which are not included in CBDB. Here we inferred 32,553 new basic kinship instances by using 6 relations: grandfather, grandmother, great-grandfather, great-grandmother, father-in-law and uncle. To be specific, the grandfather in this paper refers to the paternal grandfather unless it is clearly marked as the maternal grandfather, so does the grandmother, great grandfather and so on. Thirdly, we enriched the family networks by adding missing persons to make full use of kinship instances in CBDB, especially increase pedigree depths. We added 5,805 missing persons infer 10,337 basic kinship instances by using grandfather, great-grandfather and ancestor relations. Besides, we did detection and correction of conflicting kinship instances after each step, including instances conflicting with persons' information and instances conflicting with other instances. Finally, we generated 178,390 basic kinship instances, while finding out 3,989 inconsistencies.

By traversing all basic kinship instances, we got family trees each of which has an ancestor as the root node and generations of persons as nodes. By the above regularization in three steps, the number of family trees grows from 29,316 to 29,423. The maximum depth of a family could reach 50 generations, and the largest family has 2,112 members. It proves the effectiveness of our work on regularizing kinship instances.

In conclusion, regularizing relations named by various and complex kinship words in natural languages to the three basic relations is an effective method to construct and enrich family networks which could not be observed and counted directly in kinship databases like CBDB.

Bibliography

Shang W, Huang W. (2018). Investigating the Relationships between Scholars and Politicians in Ancient China: Taking the Yuanyou Era as an Example. Journal of the Japanese Association for Digital Humanities, 3(1): 33-48.

Dynamic social network tracking in literary texts

Van Zaanen, Menno
menno.vanzaanen@nwu.ac.za
South African Centre for Digital Language Resources, South Africa

Close and distant reading are often seen as opposites. The former is typically a fine-grained, manual analysis applied to small amounts of texts, whereas the latter performs more coarse-grained computational analyses of large amounts of texts. The close reading approach allows readers to focus on specific aspects in the text, resulting in detailed investigations. The distant reading approach, in contrast, can be used to identify large patterns, which are typically hard to identify explicitly by people due to practical limitations. Both approaches have their advantages and disadvantages. The research presented here describes a distant reading method that aims to provide information, e.g., to help direct a close reading approach.

When analyzing literature, one may be interested in the power relationships between characters in the text. There are several ways of investigating this. One of these focuses on social network analysis (e.g., Agarwal et al. (2012)), which requires the identification of a social network from the text. Ven et al. (2018) provide a simple approach that creates a social network by automatically identifying characters (using a Named Entity Recognition system) and their relationships through character co-occurrence within sentences.

One of the limitations of Ven et al. (2018)'s approach is that the social network is only created after analyzing the entire text. Relationships between the characters, however, may change throughout the text, which cannot be represented through a static social network based on the text as a whole.

Here, we provide a more flexible approach that creates social networks while allowing investigations into changes throughout the text. We first split the text in sentences, which defines the granularity of the timeline. Next, we identify all named entities (i.e., characters) in the text and keep track of which sentences they occur in (e.g., through numbering). Next, we create a finite state machine with one start and one end state. For each character, we create a path from the start to the end state with as many states as there are sentences in the text. Character markers are placed on those states in which the character occurs. To identify relationships between the characters, we merge states from different paths (which each represent locations of occurrences per character). Merging of two states marked with characters indicates that there is a relationship between these characters.

State merging takes two states in a finite state machine and replaces both by a new state. All edges going into and out of both merged states are attached to the new state. This is a common technique in the field of grammatical inference (Higuera, 2010; Heinz et al., 2015), but it has also been applied, for instance, in machine translation (Zaanen and Somers, 2005). Changing the criteria that drive the selection of states to be merged has an impact on the shape of the finite state machine after merging.

After merging states according to the selection criteria, we can count the number of (merged) states marked with multiple characters. This count represents a co-occurrence count, which describes the strength of the relationship between two characters. Performing state merging only on states with marked characters that have the same sentence number, and counting merged states throughout the entire machine, results exactly in the system of Ven et al. (2018).

The finite state machine representation has several benefits. First, we can follow the changes in the social network in the text by passing through the finite state machine from start to end node. Second, it provides more flexibility in the identification of evidence for relationships between characters. For instance, we can decide to merge states when characters occur in sentences that are close together (i.e., characters are related if they occur close together in a text, but not necessarily in the same sentence). Modifying the state merging criteria influences how the relationships between characters are measured.

Even given the added flexibility of this approach, there are still several open problems (most similar to those of Ven et al. (2018)). First, named entity recognition is not perfect, so occurrences of characters may be missed or spurious characters introduced, leading to noise in the data. Second, there are multiple ways to refer to a character, such as a first names, nick names, last names, descriptions (e.g., "the wizard"), or anaphora ("he", "she", etc.) (Bipasha, 2019). Finally, the state merging criteria have an impact on how the strength of the relationships is measured. More research is needed in these areas.

Bibliography

Agarwal, A., Corvalan, A., Jensen, J. and Rambow, O. (2012). Social network analysis of Alice in Wonderland. In, *Proceedings of the Naacl-Hlt 2012 Workshop on Computational Linguistics for Literature.*

Bipasha, T. T. (2019). Extracting social network from literary prose University of Arkansas PhD thesis.

Heinz, J., Higuera, C. de la and Zaanen, M. van (2015). *Grammatical Inference for Computational Linguistics.* Vol. 8. (Synthesis Lectures on Human Language Technologies 4). Morgan & Claypool Publishers.

Higuera, C. de la (2010). *Grammatical Inference: Learning Automata and Grammars.* Cambridge, UK: Cambridge University Press.

Ven, I. van de, Lim, C., Steenbakker, M. and Zaanen, M. van (2018). Negotiating close and distant reading: Heteroglossia and networks in Zadie Smith's White Teeth. In, *Digital Humanities Benelux, Dhbenelux.*

Zaanen, M. van and Somers, H. (2005). DEMOCRAT: Deciding between multiple outputs created by automatic translation. In, *The Tenth Machine Translation Summit, Proceedings of Conference; Phuket, Thailand.* pp. 173–80.

Project Overview: Resources and Applications for Detecting and Classifying Polarized and Hate Speech in Arabic Social Media

Zaghouani, Wajdi
wzaghouani@hbku.edu.qa
Hamad Bin Khalifa University, Qatar

Introduction

Societies are increasingly divided and polarized. This polarization is driven by two connected issues: the lack of communication between groups, and the use of hate speech. With social media speeding up the spread of hateful ideologies, polarization and technology go hand in hand. Statistics reveal the scale of the problem; 41% of people have been the target of hate speech. As communities recede into themselves, the prospect of conflict grows. Social media is also providing new opportunities for polarization and hate speech. Shielded behind anonymity, state actors and political entities are using social media to manipulate public opinion on an industrial scale, driving polarization with disinformation and hate speech to serve often extremist agendas. Combined with bots - automated accounts - these partisan entities can achieve a negative impact in society (KS Hasan, 2013 ; Howell (2013). Social media companies have been slow to tackle the problem, for instance, Facebook redefined hate speech pages as controversial humor. While Twitter introduced a new policy stating "You may not dehumanize anyone based on membership in an identifiable group, as this speech can lead to offline harm", the business model of social media companies may also not be conducive to tackling hate speech. Indeed, hate speech pages can be popular, encouraging clicks and driving advertising revenue to web companies. This is why tackling hate speech and polarization requires multilateral efforts involving the companies themselves, academics and civil society. There have been efforts to address the problem such as the efforts done by the European Commission to tackle hate speech by signing a code of conduct with social media companies to fight hate speech. However, the problem is a global issue. Despite the widespread adoption of social media in the MENA region, most efforts in tackling hate speech also tend to focus on the developed world, with little research targeting Arabic. Some of the research targetting Hate speech in Arabic were limited to a specific categories such as hate targetting religious groups (Albadi 2018) or only covering abusive language detection as in (Mubarak 2017).

Without adequate, contextual-based research, countries in the developing world in particular risk becoming social media blackspots - spaces where hate speech flourishes in unregulated and permissive online environments. The main aim of this project is to address this gap and pave the way for further research on Polarization and Hate Speech in Arab societies.

Methodology

Our research will address different problems that contribute to the detection of polarization and hate speech: 1) Stance detection with respect to controversial topics (a topic generating a polarized discussion: in favor vs. against); 2) Identification of polarized communities; 3) Hate speech detection; 4) Bot versus human identification and 5) Behavioral interventions to address hate speech. These components will be considered from a holistic perspective unlike some of the existing research works, which address them as isolated problems. Our project focuses on five components: 1) Annotated Language Resources; 2) Polarized Communities Analysis; 3) Methods and Tools based on Natural Language Processing methods as in Fersini et al. (2018); 4) Behavioral interventions and experiments to address hate speech 5) Application Scenarios

with the stakeholders. We will create annotated Arabic corpora from Twitter with the stance information (in favor, against or neutral) with respect to controversial topics (e.g., Qatar vs. UAE), polarized communities (e.g Liberals vs Conservatives) and the hateful usage of the language (e.g. insults, aggressive words). This will include creating an Arabic multi-dialectal lexicon of hate speech and aggressive language. The project has several application scenarios. In the context of cyber-security, government agencies could detect individuals and groups that spread hate speech and take appropriate countermeasures. Furthermore, bots spreading hate speech who increase the tension and polarization on society can be detected automatically. In a recent study by Jones (2016) and Jones (2019), 17% of a random sample of tweets in Arabic that mention Qatar were tweeted by bots in May 2017 and that increased to 29% in May 2018.

Conclusion

The main novelty of this proposal is in the scope, the multidisciplinary nature and the coverage of the addressed problems: we will address the main related problems to polarization as a whole, and not as isolated problems as it was done in some existing projects. Behavioral experiments and interventions will be conducted to address the issue of hate speech. We will test the state-of the-art methods of artificial intelligence to automatically approach the aforementioned problems in Arab social media. By taking into account the legal, the behavioral and the ethical dimensions of the software solutions as well as data protection considerations, we plan to create tools that will allow others to use them to detect polarization, hate speech, and bots.

Acknowledgments

This project is funded by NPRP grant NPRP13S-0206-200281 from the Qatar National Research Fund (a member of Qatar Foundation).

Bibliography

Albadi, N., Kurdi, M., & Mishra, S. (2018). Are they our brothers? analysis and detection of religious hate speech in the Arabic Twittersphere. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 69-76). IEEE. https://doi.org/10.1109/ASONAM.2018.8508247

Fersini, E., Rosso, P., Anzovino, M. (2018). Overview of the Task on Automatic Misogyny Identification at IberEval. In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018).

Hasan, K.S., Ng, V. (2013). Stance Classification of Ideological Debates: Data, Models, Features, and Constraints. In Proceeding of the Sixth International Joint Conference on Natural Language Processing

Howell, Lee. (2013). Digital Wildfires in a Hyperconnected World. WEF Report 3.

Jones, M. O. (2016). Automated sectarianism and pro-Saudi propaganda on Twitter. Exposing the Invisible (Tactical Technology Collective).

Jones, M. O. (2019). The Gulf Information War| Propaganda, Fake News, and Fake Trends: The Weaponization of Twitter Bots in the Gulf Crisis. International Journal of Communication

Mubarak, H., Darwish, K., & Magdy, W. (2017). Abusive language detection on Arabic social media. In Proceedings of the first workshop on abusive language online (pp. 52-56).

How to Critically Utilise Wikidata – A Systematic Review of Wikidata in DH Projects

Zhao, Fudie
fudie.zhao@sant.ox.ac.uk
University of Oxford, United Kingdom

Initiated in 2013, Wikidata is a free and open knowledge base that acts as central storage for the structured data of its Wikimedia sister projects. It has been adopted and systematically reviewed in Information Science/Computer Science (Mora-Cantallops et al., 2019) and the library domain (Tharani, 2021). Projects in the DH domain have also been embracing Wikidata in their data-related activities. For example, since 2016, 43 presentations at DH conferences held by ADHO have mentioned Wikidata in their abstracts, as shown in **Fig.1**. [1] However, except Stacey's paper about Wikidata's use in GLAMs and DH (Stacey, 2017), there still lacks a systematic review regarding Wikidata's status quo, potential, and challenges in the field.

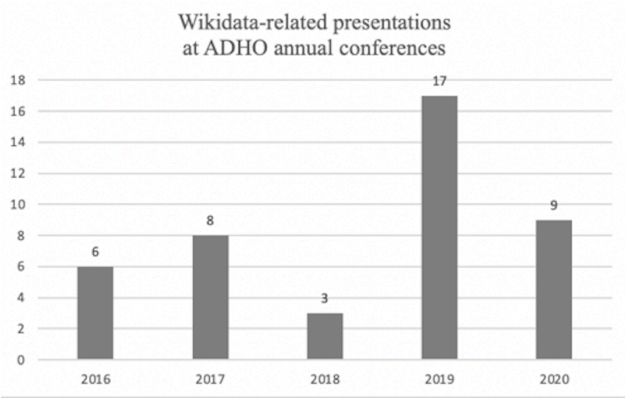


Fig.1:
Wikidata-related presentations at ADHO annual conferences

This short paper intends to fill this research gap by proposing four research questions:

Q1: How is Wikidata described in the current DH literature?

Q2: To what end is Wikidata being experimented within the DH domain?

Q3: What are the potentials of embracing Wikidata in data-related activities in DH projects?

Q4: What are the challenges and possible solutions associated with Wikidata in DH projects?

To answer the questions, a systematic literature review of DH projects that adopted Wikidata has been conducted based on the guidelines for Systematic Review proposed by Kitchenham (2004). Book of Abstracts from ADHO annual conferences, a compiled list of DH journals, and five online academic research databases (ACM Digital Library, Springer Link, and Web of Science, Science Direct) were searched and screened, guided by pre-determined search strategies and inclusion & exclusion criteria. **196** papers/presentations were identified in the sources, and after the screening, **58** were selected based on criteria (English only, no duplicates, only application studies, Wikidata implemented) for further analysis in **Table 1**. [2]

Sources	Search strategy	No. of articles and presentations identified
ADHO annual conferences	'Wikidata' in title, abstract, keywords or text	43
DH journals	'Wikidata' in title, keywords or text	24
Wikimedia platform (Diff)	'digital humanities' in title or text	1
ACM Digital Library	'Wikidata' AND 'digital humanities' in title, abstract, keywords, or text	22
IEEE Xplore		0
Springer Link		82
Science Direct		7
ISI Web of Science		17

Table 1:
Total number of articles and presentations identified from each source

This paper finds that:

The descriptions of Wikidata in the current DH literature fall into three categories: a **technology stack** to access Linked Data, a **platform** for crowdsourcing, collaboration, dissemination, and linking datasets on the Semantic Web, and a **content provider** of open, free, generic, editable, heterogeneous, linked data, as shown in **Fig.2**:

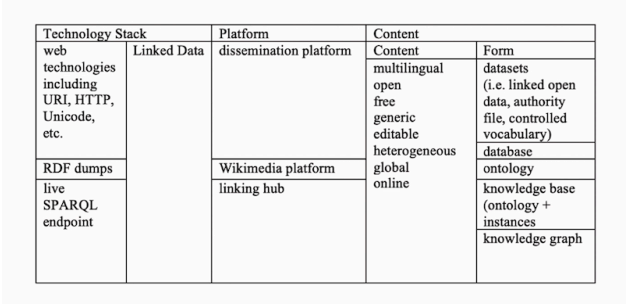


Fig. 2:
Wikidata Components

Wikidata has been included in data-related tasks such as annotation and enrichment, metadata curation, named entity recognition and disambiguation, knowledge representation and ontological engineering, data sourcing, aggregation of datasets, and the pursuit of open citation data and pedagogical practices (miscellaneous) as shown in **Table 2**.

	Occurrence frequency	
	Count	Percentage
semantic annotation and enrichment	22	37.9%
metadata curation	14	24.1%
named entity recognition, disambiguation, linking	7	12.1%
knowledge representation and ontological modelling	5	8.6%
data sourcing	4	6.9%
data aggregation	3	5.2%
miscellaneous	3	5.2%
total	58	100.0%

Table 2:
Wikidata application areas in the reviewed items

Projects in the DH domain can use Wikidata for data consumption and publication:

1) Data consumption – Wikidata is a data source for enrichment.

2) Data publication and exchange – Wikidata is an access point to disseminate data to the broader landscape of the Web for public engagement; a platform for crowdsourcing and collaborative production of linked data; a linked data approach towards the integration of data within a specific domain.

The use of Wikidata is accompanied by doubt about its data quality. Cook (Cook, 2017, 122) points out that Wikidata's data is too generic and short of quality for DH scholars who tend to work in a specific area, while Wikimedians pay less attention to research-oriented DH

projects and focus more on projects which gather data and edit pages. The DH community can learn from the technical community regarding the factors that influence its data quality, and possible solutions. Factors specified in the research include: user types and their editing activities, the effectiveness of systems and tools to facilitate detection and improvement of data quality, and the relevance and authoritativeness of its external references and sources. The solutions proposed by the technical community encompass 1) a better understanding of users and the editorial process via research, and 2) the development of systems, measures, and tools concerning the evaluation and improvement of different dimensions of data quality. The technical side, however, has its limitation. As pointed out by the IS systematic review (Mora-Cantallops et al., 2019, 262), such applications are mostly limited to Wikidata itself and are yet to be linked to disciplines outside information systems. The contribution of this paper is to address three factors and relevant solutions in the specific context of DH projects: the relevance and authoritativeness of other available domain sources, domain communities and their activities, and workflow designs that balance the automated and manual work by utilising the technical and labour resources of a project's own and those offered by Wikidata.

This paper intends to invite discussion from participants at DH2022 about Wikidata's possible use in the DH context and the challenges it may face.

Bibliography

Cook, S. (2017). The uses of Wikidata for galleries, libraries, archives and museums and its place in the digital humanities. *Comma*, **2017**(2):117-124.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, **33**(2004), 1-26.

Mora-Cantallops, M., Sánchez-Alonso, S. and García-Barriocanal, E. (2019). A systematic literature review on Wikidata. *Data Technologies and Applications*, **53**(3): 250–68.

Tharani, K. (2021). Much more than a mere technology: A systematic review of Wikidata in libraries. *The Journal of Academic Librarianship*, **47**(2).

Notes

1. Data about the ADHO annual conferences is collected from *the Index of Digital Humanities Conferences* site which aggregates and presents conference metadata: https://dh-abstracts.library.cmu.edu/conferences
2. Until December 31, 2021.

Multimedia Retrieval of Historical Materials

Zhu, Jieyong
zjsczjy04@gmail.com
Graduate School of Informatics, Kyoto University

Nishimura, Taichi
taichitary@gmail.com
Graduate School of Informatics, Kyoto University

Goto, Makoto
m-goto@rekihaku.ac.jp
National Museum of Japanese History

Mori, Shinsuke
forest@i.kyoto-u.ac.jp
Academic Center for Computing and Media Studies, Kyoto University

Introduction

Historical material is a collection of history, archaeology, and folklore materials. With the rapid advancement of digitization, large-scale multimedia data of historical materials have become available on the web. As the data grows, it is difficult for researchers to study the relationship between historical images and text. Multimedia retrieval is a technique to perform retrieval tasks across multiple media. Recently, deep learning has accelerated research on natural language understanding and computer vision, with remarkable performance reported in multimedia retrieval tasks (Salvador et al., 2017). In this paper, we apply the state-of-the-art multimedia retrieval methods to Japanese historical materials and demonstrate the constructed multimedia retrieval system. Fig 1 shows an example of multimodal retrieval tasks of historical materials.

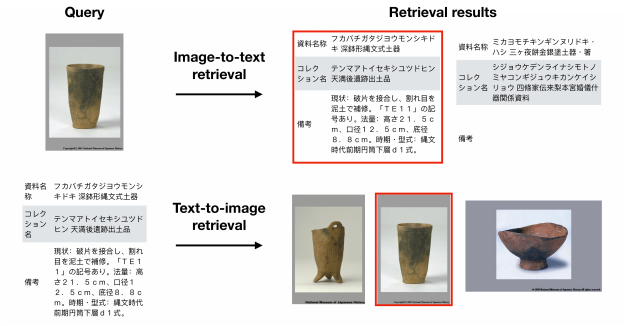


Fig 1.
Examples of image-to-text and text-to-image retrieval tasks. Retrieved results in the boxes are the associated ones with the query on the left.

Multimedia Retrieval

Multimedia retrieval takes one type of media (e.g., images and texts) as the query to retrieve corresponding media of another type (Liu et al., 2010). The key challenge of multimedia retrieval is how to convert different media data into a shared subspace, where semantically associated inputs are mapped to similar locations. Various kinds of deep-learning-based approaches have been proposed in the literature (Sirirattanapol et al., 2017). We here employ one of them to realize our system.

Proposal

This paper proposes a deep-learning-based approach to achieve multimedia retrieval of historical materials. Figure 2 shows an overview of our proposed model. The proposed method consists of two major processes.

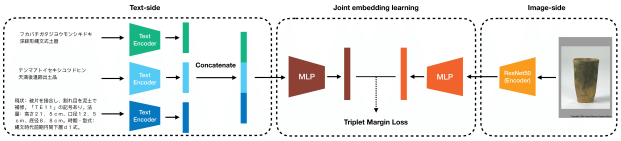


Fig 2.
An overview of our proposed method.

Text encoder

Recently, large-scale pre-trained model, such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), has achieved great performance in NLP tasks. However, we don't use BERT model because of the domain gap, since Japanese BERT is trained on Japanese Wikipedia texts. The text data of historical materials include three main types: material name, collection name, and notes. All the three types of data are in a tabular format instead of a sentence format. Therefore, we use a word2vec model to convert the texts into vectors, which is more simple and more reasonable.

Image encoder

The input of the image side is a single image of historical materials. To convert images into vectors, we employ ResNet50 (He et al., 2016), a Convolutional Neural Network pre-trained on ImageNet.

Shared subspace learning

Finally, we convert text/image vectors into shared subspace using symmetric multi-layer perceptrons with ReLU activation functions. To train the model, we compute triplet margin loss (Vassileios Balntas and Mikolajczyk, 2016), which makes the vectors in the subspace for a given text-image pair close and otherwise long.

Dataset

This is the first attempt to tackle multimedia retrieval of historical materials, so no datasets exist in this field; thus we created the Japanese historical dataset of textual descriptions and corresponding images by crawling them from the National Museum of Japanese History. The dataset contains 18,429 objects, including over 18k textual descriptions and over 79k corresponding images.

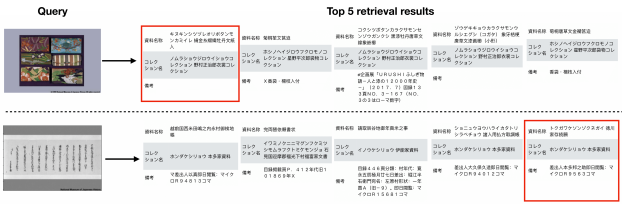


Fig 3.
Image-to-text retrieval examples. The ground truth in the retrieved results is highlighted in the box.

Experiments

To measure the performance of the model, we perform multimedia retrieval tasks. Figure 3 shows two examples of the image-to-text task. The query images are on the left side while the top five retrieved texts are on the right side. As with previous studies, we compute three mainstream

evaluation metrics, median rank (MedR), Recall@K (R@K) (Salvador et al., 2017), and mean average precision (mAP) (Rasiwasia et al., 2010) to evaluate the performance. Table 1 shows the results of 1,000 samples. The result indicates that our system performs well in multimedia retrieval tasks compared with the random ranking baseline.

	Image => Text	Text => Image	Random Ranking
R@1	0.036	0.043	0.001
R@5	0.144	0.163	0.005
R@10	0.258	0.285	0.01
medR	26	26	500
mAP	0.107	0.119	0.002

Table 1.
Retrieval results on 1,000 samples.

Conclusion

This paper tackled the multimedia retrieval of historical materials using deep-learning-based multimedia retrieval methods. This work is the first attempt to tackle this problem, thus we constructed the dataset of Japanese historical texts and images, and evaluated the model's performance on it. The experimental results show that our constructed system performs well in the multimedia retrieval of historical materials. Future work will study a better method to represent the textual data. We expect that our research will help researchers in gaining a better understanding of Japanese historical materials, and will give a general approach to learning the shared subspace between textual and visual data.

Bibliography

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–86 doi:10.18653/v1/N19-1423. https://aclanthology.org/N19-1423.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 770–78.

Liu, J., Xu, C. and Lu, H. (2010). Cross-media retrieval: state-of-the-art and open issues. *International Journal of Multimedia Intelligence and Security*, **1**(1). Inderscience Publishers: 33–52.

Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G. R., Levy, R. and Vasconcelos, N. (2010). A new approach to cross-modal multimedia retrieval. *Proceedings of the 18th ACM International Conference on Multimedia*. pp. 251–60.

Salvador, A., Hynes, N., Aytar, Y., Marin, J., Ofli, F., Weber, I. and Torralba, A. (2017). Learning cross-modal embeddings for cooking recipes and food images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3020–28.

Sirirattanapol, C., Matsui, Y., Satoh, S., Matsuda, K. and Yamamoto, K. (2017). Deep image retrieval applied on kotenseki ancient japanese literature. *2017 IEEE International Symposium on Multimedia (ISM)*. IEEE, pp. 495–99.

Vassileios Balntas, D. P., Edgar Riba and Mikolajczyk, K. (2016). Learning local feature descriptors with triplets and shallow convolutional neural networks. In Richard C. Wilson, E. R. H. and Smith, W. A. P. (eds), *Proceedings of the British Machine Vision Conference (BMVC)*. BMVA Press, p. 119.1-119.11 doi:10.5244/C.30.119. https://dx.doi.org/10.5244/C.30.119.

Electronic Posters

The Pianolatron: Enabling Web-Based Interactive Performances of Digitized Player Piano Rolls

Abraham, Vijoy
vijoy@stanford.edu
Stanford University, United States of America

Arul, Kumaran
karul2@stanford.edu
Stanford University, United States of America

Barth, George
barth@stanford.edu
Stanford University, United States of America

Broadwell, Peter
broadwell@stanford.edu
Stanford University, United States of America

Wiles, Simon
simon.wiles@stanford.edu
Stanford University, United States of America

The Pianolatron (pianolatron.stanford.edu) is a web browser application developed by library-based researchers in collaboration with faculty principal investigators from the Department of Music. The project's objective is to make the world's largest library-based collection of digitized historical player piano rolls more fully accessible to scholars and the general public. The app plays a digitized roll based on data from an image pre-processing step that detects the positions of the punch holes on the roll and converts them into MIDI events, which the app then synthesizes into a musical performance in real time via high-fidelity piano sound samples. IIIF protocols facilitate synchronization of the audio playback with a scrolling animation of the high-resolution roll image served from a digital repository. Supplementary animated graphics of a piano keyboard and pedals as well as color-coded overlay visualizations on the rolls illustrate the effect of each perforation on the music that is heard during playback.

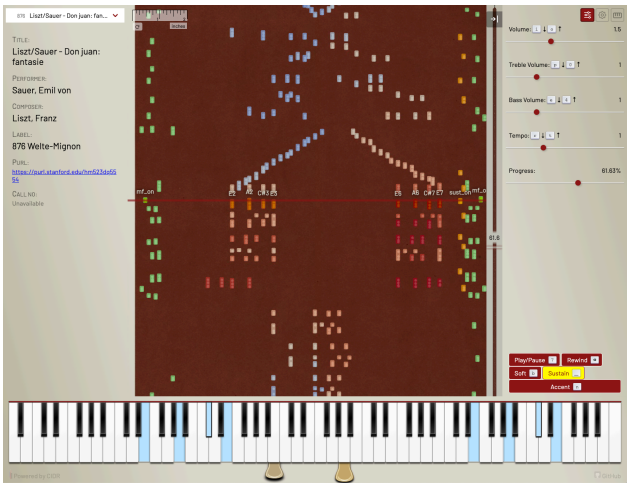


Figure 1.
The Pianolatron app playing a reproducing piano roll: the punch holes controlling dynamics and pedaling on the left and right margins of the roll are highlighted in green and orange, respectively, while the note perforations in the center of the roll are highlighted on a color scale to indicate the velocity with which the keys were struck (blue = soft, red = loud)

"Reproducing" player piano rolls, which encode the expressive nuances of a live performance, provide the most detailed remaining evidence of the performing practices of pianists and composers who were active in the nineteenth century; they were only superseded by advances in electronic sound recording and playback in the late 1920s. Rolls for the more common "pianola" player pianos, which could be found on every continent by the early twentieth century, enabled interactive performance of programmed piano tunes on a home console and are a key source of information about the popular tastes of the era.

The academic and hobbyist communities who study piano rolls and players have produced other software tools to convert scanned piano roll images into playable MIDI files and to visualize some aspects of the roll playback process for certain rolls, but the Pianolatron is the first application able to run in a web browser on any modern computer or large tablet that can play back any scanned roll in synchrony with its image. Furthermore, the application provides controls for on-the-fly modification of roll speed (tempo), volume levels and pedaling, enabling interactive performance of a roll in the same manner as an early twentieth-century expert "pianolist"—a role and activity that was roughly equivalent in popularity and effect to modern DJing. For reproducing rolls, the app exposes even more fine-grained options to allow manipulation of the parameters of the physical/pneumatic expression systems originally used to encode and replay the musical details of a live performance. The workings of many of the systems

that player piano companies developed to record in-studio performances by famous pianists for reproduction on piano rolls have been lost to history, so the ability to experiment with different settings within the app and immediately to hear (and see) their effects on the roll playback represents a major contribution to ongoing research into these topics.

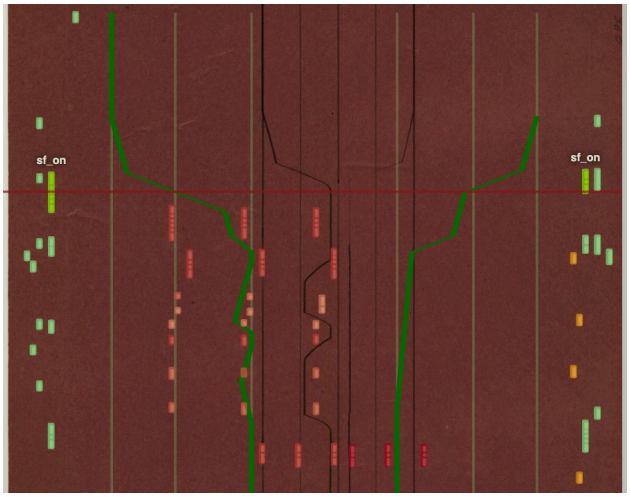


Figure 2.
A detailed view of the optional research-oriented visualizations the Pianolatron can display on a reproducing roll: the dark green curves represent the bass and treble velocity (volume) levels as calculated according to the expression punch holes that are highlighted in green on the left and right margins of the roll; these generally align well with the black "hand nuance" dynamics curves seen in the middle of the image, which were stenciled onto the roll when it was first printed in 1905

The Pianolatron player app currently enables interactive playback of nearly 1,000 piano rolls, and this number will continue to increase as new rolls are scanned and added to the collection. The use of IIIF to access the roll images also means that the app potentially can play and visualize scanned piano rolls from any repository worldwide that makes its images available via IIIF. Furthermore, the modular design of the application's source code and data files, which are fully available on Github (github.com/sul-cidr/pianolatron), encourages contributions from external researchers.

The Pianolatron application is written in the Svelte Javascript framework, which provides the modularity and support for multiple independent interactive components offered by popular frameworks such as React, Vue and Angular, while placing fewer computational demands on the web browser.

Bibliography

Colmenares, G., et al. (2011). Computational Modeling of Reproducing Piano Rolls. *Computer Music Journal* 35(1): 58–75.

Nettheim, N. (2019). Overcoming Piano-Roll Limitations: Pachmann Plays Sabouroff's Polka. *Music Performance Research* 9: 128-161.

Phillips, P. (2017). *Piano Rolls and Contemporary Player Pianos: The Catalogues, Technologies, Archiving and Accessibility*. Ph.D. thesis, University of Sydney.

Shi, Z., Arul, K., Smith, J. (2017). Modeling and Digitizing Reproducing Piano Rolls. *The 18th International Society for Music Information Retrieval Conference (ISMIR)*. National University of Singapore, pp. 197-203.

Shi, Z., et al. (2019). SUPRA: Digitizing the Stanford University Piano Roll Archive. *The 20th International Society for Music Information Retrieval Conference (ISMIR)*. Delft, Netherlands, pp. 517-523.

Digital City: Developing Digital Guerilla Tactics for the Urban Environment

Altin, Ersin
ersin.altin@njit.edu
New Jersey Institute of Technology, United States of America

Guerilla Studio is a collaborative design studio developed for senior design (interior, industrial and digital design) students and offers a number of novel pedagogies and teaching methodologies. Guerilla Studio is designed to explore possible ways to communicate a comprehensive and inclusive notion of city through design, by utilizing a number of digital tools such as VR and AR technologies, data processing, and digital visualization. While integrating digital tools to the pedagogy of design, Guerilla Studio motivates students to think about social responsibilities by widening the area that design serves and strategizing design processes as social tasks.

To determine a strategy for their design process students are asked to adopt a number of approaches such as preparing interactive "psychogeographic" maps of a portion of a city after experiencing it by adopting Situationist techniques of dérive and detournement. Another exercise mediates non-human dwellers of cities by utilizing data processing and visualization tools. Following a number of design exercises, students develop a wide range of design tactics varying from adding elements of (interior, industrial and digital) design to existing structures to "hack"

a certain dominant narrative to more literal approaches which propose physical or digital "design solutions" that communicate the collective alternative information about the/any city.

Guerilla Studio locates each design problem within the urban context. In doing so, design itself and the cultural problem it tackles become a ground to discuss larger issues and, conversely, those societal/cultural problems gain a physicality within urban space. The goal is to create a correlation between micro and macro scales by stressing the complex and dynamic connection between them. In this sense, any intervention to urban space with design would become the outcome of various analytic phases in design process such as literature review, investigation of emerging technologies, analysis of urban space by concentrating on smaller scale. More importantly, design is recognized to have the potential to trigger reaction by becoming a part of urban life rather than isolated instances of production.

One of the student works produced in Guerilla Studio, for instance, aims to confront modern city dwellers and their dependency on mobile devices by claiming their mobile phones temporarily. While mobile technology is undeniably one of the greatest tools currently available to mediate information, help, or leisure to list a few, its increasing ubiquity and demand for attention can create a barrier between the users and their overall experience with their surroundings. This project invites participants to engage with others and with the urban environment by plugging into the urban digital network without interruptions if they agree to give up their mobile device for a short time.

Another student work explores the physicality of urban space and related cultural patterns as a research tool in heritage studies. The project models and visualizes the city of Chan Chan by utilizing VR technologies to introduce this important ancient American city to the inhabitants of contemporary American cities. While a portion of Chan Chan can be experienced through a VR headset remotely, the project also seeks to find ways to create intersections between the culture of Chan Chan and present cities/city dwellers by organizing Guerilla shows in the mainstream art and culture centers, such as the Metropolitan Museum of Art.

Guerilla Studio encourages students to create a system that functions as informal nodes of information that are neither located nor created centrally but as part of the larger network of communications/interactions by exploring limits of digital tools and utilizing them creatively. Guerilla Studio points out various possible intersections/interactions between the physicality of urban fabric and more abstract and complex domains such as digital technologies, culture, and identity.

Bibliography

Debord, G. (1977). *Society of the Spectacle.* Detroit: Black and Red.

Moseley, M.E. and Mackey, C.J. (1974). *Twenty-Four Architectural Plans of Chan Chan, Peru: Structure and Form at the Capital of Chimor.* Cambridge: Peabody Museum Press.

Topic, J.R. (2003). From Stewards to Bureaucrats: Architecture and Information Flow at Chan Chan, Peru. *Latin American Antiquity* 14(3): 243–74, https://doi.org/10.2307/3557559.

Revitalizing a South Asian language with Unicode: The case of Sunuwar in Nepal

Anderson, Deborah {Debbie}
dwanders@sonic.net
UC Berkeley, United States of America

Sunuwar, Dev Kumar
devkumarmail@gmail.com
Indigenous Media Foundation, Nepal

Background

The Sunuwar, or Koĩts-Lo, language is spoken by 37,900 users (2011 census) (Eberhard et al., 2021) in Nepal, and is also spoken in Sikkim, India. The Sunuwar language is classified by the EGIDS scale as a "threatened" language, indicating the language is losing speakers (SIL International, n.d.).

This poster describes ongoing work to revitalize the Sunuwar language, using Unicode, as a possible model for other indigenous languages. It will include examples of the script, with sound bites of the language and discuss the steps in getting a script into Unicode.

Although Sunuwar is primarily oral language, a script called "kõits brese" was devised for the language in the 1940s. This script was promoted in the 1960s and 1970s (Sunuwar, 2021), but these efforts were hindered by Nepal's one language policy: *ek bhasha, ek bhesh, ek dharma, ek desh* ("one language, one way of dress, one religion, one nation"). This "one language" policy implied the use of the Devanagari script, which is used to write the official Nepali language. The policy lasted until 1990 (Weinberg, 2013).

In the 2000s, the Government of Nepal developed various plans that promoted mother tongue multilingual education. Currently, 24 school curriculums in different languages have been created, but the medium for these

materials for different languages -- including Sunuwar -- is in the Devanagari script (Sunuwar, 2021). This situation is due in part to the history of language policies in Nepal, which did not encourage use of scripts other than Devanagari. Another key factor is that Sunuwar, as well as several other scripts of smaller language communities in Nepal, are not in the Unicode Standard, which means creating and exchanging text in the script electronically is very difficult.

Digitization Process and Unicode

The Sunuwar Welfare Society, which was established in 1988, has long envisioned having the language move from an oral language to one whose written version can be sent and received on various digital platforms. Digitization will help to preserve the language and its script and strengthening the distinct Sunuwar identity.

In 2020, the Translation Commons project, a partner with UNESCO's International Year of Indigenous Languages and International Decade of Indigenous Languages, selected Sunuwar as a pilot project to demonstrate the scalability of encoding scripts for indigenous languages and has facilitated the Sunuwar encoding project.

A key first step towards digitization was to get the script into the Unicode Standard. Fortunately, a Unicode proposal had earlier been written in 2011 by Anshuman Pandey (Pandey, 2011). The proposal was reviewed in 2020 by a core team of linguists, educators, activists, journalists, and language practitioners who are members of the Sunuwar Welfare Society (Sunuwar, 2021). Regular meetings took place between Dev Kumar Sunuwar (Indigenous Media Foundation), Anshuman Pandey and Deborah Anderson (Script Encoding Initiative), Craig Cornelius (Google), and Jeannette Stewart (Translation Commons) which ironed out questions for the Unicode proposal, a font, and keyboard. At the January 2022 UTC meeting the proposal was approved for inclusion in a future version of the Unicode Standard.

Challenges

To be approved for inclusion in Unicode, evidence showing usage of the script is needed.

Being primarily an oral language used widely in the home, Sunuwar does not have an extensive written tradition nor has it been widely used in Nepal. However, in Sikkim, India, the Sunuwar language (called "Mukhia") was officially recognized in in 1996, and the Sikkim government published schoolbooks, newspapers and other materials in the script (Sunuwar, 2021). The evidence from Sikkim is useful for demonstrating usage, but evidence from Nepal would be needed to confirm consistent usage across Nepal and India.

Recent promising signs of Sunuwar script usage include:

- Creation of a keyboard and PUA-based font, which can be used to develop more written materials.
- Appearance in 2021 of a monthly magazine *Hamso* in Sunuwar and Devanagari scripts (in the Sunuwar language) (Sunuwar, 2022).
- Creation of YouTube instructional videos and in-person classes on the script, and an alphabet book.

Future

This project could serve as a model for other communities in Nepal (and beyond) that can learn from the Sunuwar experience on how to get a script into Unicode and revitalization efforts, including any difficulties encountered. More generally, what can be learned from the Sunuwars' experience of transitioning from an oral culture to a written one that might be applicable to other communities? How can other primarily oral communities build up their written culture, and thus have their script be eligible for inclusion in the Unicode Standard?

The result of the work to get the script into Unicode and widely adopted is yet to be seen. The predominance of Nepali in Nepal and English for business and international communication has tended to overshadow lesser-used languages. For example, today, English is used as medium of instruction in some private schools in Nepal and may be adopted into public schools as well (Phyakm, 2021).

Funding

This work was supported by National Endowment for the Humanities [grant PR-268710-20].

Bibliography

Eberhard, David M., Gary F. Simons and Charles D. Fennig (eds.). (2021). *Ethnologue: Languages of the World*. Twenty-fourth edition. Dallas, Texas: SIL International, online edition.

Pandey, Anshuman. (2011). "Proposal to Encode the Jenticha Script." https://www.unicode.org/L2/L2011/11218-n4028-jenticha.pdf (accessed 6 April 2022). (Note: The proposal used the script name "Jenticha," but this has subsequently been changed to "Sunuwar.")

Phyak, Phrem. (2021). Language education policy in Nepal and the denial of the right to speak in Indigenous Languages. *Melbourne Asia Review*, Edition 7, 2021. https://melbourneasiareview.edu.au/language-education-policy-in-nepal-and-the-denial-of-the-right-to-speak-in-indigenous-languages/ (accessed 6 April 2022).

SIL International (no date). *Language Status*. https://www.ethnologue.com/about/language-status (accessed 6 April 2022).

Sunuwar, Dev Kumar. (2021). "Digitizing the script of Koits Sunuwar Indigenous Peoples." https://

www.devkumarsunuwar.com.np/digitizing-the-script-of-koits-sunuwar-indigenous-peoples (accessed 6 April 2022).

Sunuwar, Dev Kumar. (2022). *Hamso* magazine. https://www.devkumarsunuwar.com.np/hamso-magazine (accessed 6 April 2022).

Weinberg, Miranda. (2013). "Revisiting History in Language Policy: The Case of Medium of Instruction in Nepal." *Working Papers in Educational Linguistics* 28 (1): 61-80. https://repository.upenn.edu/wpel/vol28/iss1/6 (accessed 6 April 2022).

RELEVEN: Re-evaluating the Eleventh Century through Linked Events and Entities

Andrews, Tara Lee
tara.andrews@univie.ac.at
Institute for History, University of Vienna, Austria; Austrian Centre for Digital Humanities and Cultural Heritage, Austrian Academy of Sciences

Ebel, Carla
carla.ebel@oeaw.ac.at
Austrian Centre for Digital Humanities and Cultural Heritage, Austrian Academy of Sciences

Richards, Nina
nina.brundke@oeaw.ac.at
Austrian Centre for Digital Humanities and Cultural Heritage, Austrian Academy of Sciences

Prajda, Katalin
katalin.prajda@univie.ac.at
Institute for History, University of Vienna, Austria

Rózsa, Márton
marton.rozsa@univie.ac.at
Institute for History, University of Vienna, Austria

Anđelović, Aleksandar
aleksandar.andjelovic@univie.ac.at
Institute for History, University of Vienna, Austria

Read, Lewis
lewis.read@univie.ac.at
Institute for History, University of Vienna, Austria

Introduction

This poster presents the work of RELEVEN, an ERC-funded project running at the University of Vienna from 2021–2026. The aim of RELEVEN is to cast a clearer light on the events of the "short eleventh century" (c. 1030–1095) and specifically to employ digital means of historical information modelling to seek a more coherent and connected picture of this period, to match the coherent and increasingly connected Christian world that Shepard (2017) and Frankopan (2013), among others, have argued for. We seek to understand how the different people in this world conceived of it in terms of the space they inhabited, the people around them, and the written artefacts, pointing in many cases to the intellectual ideas that circulated around it. The importance of such a connected approach is clear in light of recent trends toward thinking in terms of a Global Middle Ages (cf. Holmes and Standen 2018). The key to achieving our aim is to find a way to link and connect large amounts of disparate sorts of data. We aim to find a model for expressing data about the eleventh century that allows us to incorporate and model different, and even conflicting, perspectives about what the data tell us.

Methodology

If digital data is to be useful for historians, it must be directly linkable not only to provenance in the sense of primary source material, but primarily to the authority of the scholar who is interpreting the primary source(s) to make the claim. This principle is implemented in our STAR (Structured Assertion Record) data model, which for historical information about people and places is based on existing standards such as CIDOC-CRM (Bekiari et al. 2021) and the Linked Places specification (Grossner 2016; 2022), used together with the 'Proxy' concept of the OAI-ORE data model (Lagoze et al. 2008) when we need to instantiate multiple competing versions of events. Both existing and new historical data is represented as sets of assertions along these lines, often sourced but always linked to an authority; this allows data to be manipulated according to source and authority, and also allows assertions themselves to be linked depending on whether they corroborate, depend on, or conflict with each other. Movements of people and objects can be mapped according to different reconstructions; the interchange of ideas between people and groups can be drawn, or re-drawn, in competing schematics according to the ideas of different scholars. The novel aspect of this methodology is that it takes to its logical conclusion something that historians all readily acknowledge and that is especially apparent for pre-

modern history: that there are very few, if any, simple and undisputed facts.

Trans-regional approach

Our approach is tested by taking a broad trans-regional approach to the history of the late 11th century (c. 1030–1095), centred broadly in the eastern half of Christendom; our project focuses in particular on Byzantium and the Caucasus, on buffer zones between West and East such as Istria and the newly-established Kingdom of Hungary, and on the process of Christianisation in central and northern Europe. The looming weight of the First Crusade at the century's end means that while certain regional or proto-national narratives—particularly for western Europe—are well-developed, they tend to obscure the larger trans-regional trends of communication and contact, particularly in eastern Christendom.

By drawing upon the depth of scholarship and the plethora of digital resources that have emerged for this period in sub-disciplines such as prosopography, textual scholarship, corpus-based research, and archaeology, and by framing this scholarship in terms of assertions whose authority is traceable, we aim in this project to look at the history not just from "the eastern perspective", but from several perspectives at once.

Acknowledgments

The project described here has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 101002357).

Bibliography

Bekiari, C., Bruseker, G., Doerr, M., Ore, C.-E., Stead, S. and Velios, A. (eds). (2021). Definition of the CIDOC Conceptual Reference Model https://www.cidoc-crm.org/Version/version-7.2.

Frankopan, P. (2013). *The First Crusade: The Call from the East*. Random House.

Grossner, K. (2016). Linking Linked Places *Kgeographer* http://kgeographer.org/linking-linked-places/ (accessed 10 December 2021).

Grossner, K. (2022). *LinkedPasts/Linked-Places-Format*. Linked Pasts https://github.com/LinkedPasts/linked-places-format (accessed 6 April 2022).

Holmes, C. and Standen, N. (2018). Introduction: Towards a Global Middle Ages. *Past & Present*, **238**(suppl_13): 1–44 doi: 10.1093/pastj/gty030.

Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R. and Warner, S. (eds). (2008). ORE Specification - Abstract Data Model https://www.openarchives.org/ore/1.0/datamodel (accessed 10 December 2021).

Shepard, J. (2017). Storm clouds and a thunderclap: East-west tensions towards the mid-eleventh century. In Whittow, M. and Lauxtermann, M. D. (eds), *Byzantium in the Eleventh Century: Being in Between*. Abingdon: Routledge, pp. 127–53.

Visualization of annotation system of the Theravada Buddhist literature

Aono, Michihiko
aonomichihiko@hotmail.com
International Institute for DigitalHumanities, Japan

The Theravada Buddhist texts consists of a huge number of scriptures and commentaries. The main parts of them are handed down in one of the Middle Indo-Aryan languages called Pali or Magadhi. So far, they have occupied the most important position in the study of Buddhist literature, since they can be read more strictly than Chinese translations, Sanskrit manuscripts and so on.

Around the turn of the millennium, the Vipassana Research Institute (India), the Dhammakaya Foundation (Thailand) and Mahidol University (Thailand) released a voluminous amount of the digitized texts of the Theravada Buddhist texts written in Pali language. Now that it is possible to read them on personal computers and perform full-text search, most of Buddhologists use the digitized texts for their research instead of printed books and concordances.

However, I wonder if this can be called "digital shift". We are only reading the digital texts on personal computers, without digitalizing our research at all. Our research method is still the same as before. In order to renew our philological study of the Theravada Buddhist texts and to achieve the digital shift in the true sense, I feel certain that it is essential to structuralize the digital texts. This is the reason why I started to mark up the structure of the scriptures and their commentaries in accordance with Text Encoding Initiative P5 Guideline.

In this poster presentation, I will point out some problems which I encountered when trying to apply the Text Encoding Initiative P5 Guideline to the complicated

structure of the annotation system of the Theravada Buddhist texts. To give an example concisely here, <gloss> may not be available to structuralize the annotation system. This is because the commentaries of the Theravada Buddhist texts sometimes cite sentences from other sources or include popular verses. In these cases, it is necessary to use <quote> indicating a quotation and <l> and <lg> indicating a verse and a group of verses. However, <gloss> cannot include <quote>, <l> and <lg> according to the Text Encoding Initiative P5 Guideline. In the first place, "gloss" (Gk. Glossa, Lat. Glossa) is a simple note or comment added to a piece of writing to explain a difficult word or phrase in the line spacing and margins, based on the definition of *A Dictionary of the English Bible and Its Origins*. While "Gloss" is dependent on the main text, the commentaries of the Theravada Buddhist texts are independent of the main texts, i.e. the scriptures. In order to overcome these difficulties and structuralize the annotation system accurately, I decided to adopt <note> with @type as an alternative to <gloss>. As in the above example, I will show the other difficulties of encoding and how to solve them in accordance with Text Encoding Initiative P5 Guideline in this presentation.

Bibliography

Gilmore, Alec. (2000). *A Dictionary of the English Bible and Its Origins*. Sheffield: Sheffield Academic Press.

TEI Consortium. (2021). TEI P5: Guidelines for Electronic Text Encoding and Interchange 4.3.0, https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf (accessed 31 March 2022).

A Study on the Accuracy of OCR-based and NLP-based detection of Japanese Text in the HathiTrust Extracted Features v2.0 Dataset

Bainbridge, David
davidb@waikato.ac.nz
Department of Computer Science, University of Waikato, New Zealand

Hilbing, Genna
hilbing4@illinois.edu
HathiTrust Research Center, iSchool, University of Illinois Urbana-Champaign, USA

Jiang, Ming
mjiang17@illinois.edu
HathiTrust Research Center, iSchool, University of Illinois Urbana-Champaign, USA

Hu, Yuerong
yuerong2@illinois.edu
HathiTrust Research Center, iSchool, University of Illinois Urbana-Champaign, USA

Layne-Worthey, Glen
gworthey@illinois.edu
HathiTrust Research Center, iSchool, University of Illinois Urbana-Champaign, USA

Downie, J Stephen
jdownie@illinois.edu
HathiTrust Research Center, iSchool, University of Illinois Urbana-Champaign, USA

The HathiTrust Research Center (HTRC) Extracted Features (EF) Dataset [1] consists of volume-, page-, and word-level data for more than 17 million volumes in a wide array of languages. Every volume is described by a library catalogue record, which includes at least one cataloguer-determined primary language for that volume. Although it is generally accurate, volume-level language information does not tell the whole story of a book: such description likely disregards substantial but incidental additional language material at the page level. Accompanying and supplementing this human-created, volume-level language metadata in the HTRC EF Dataset is page-level, machine-generated language metadata for each of the 6.2 billion pages—a design decision we consider appropriate, given the overwhelmingly daunting task that page-level manual cataloguing would be.

Machine-generated language detection occurs at two different stages of the EF production process: during initial OCR, and as part of a complex pipeline of other natural language processes [5, 6]. This poster reports on a set of related studies to assess the quality and usability of this machine-generated metadata, and to suggest means to improve them. In recognition of DH2022's host country, and acknowledging that both NLP and OCR are notoriously problematic for Asian languages [2, 3, 4], we have narrowed our focus here on texts identified by either human or algorithm as being in Japanese.

Both page-level and volume-level metadata are searchable in HTRC's Solr-based search interface, the "Workset Builder," which, in an ideal scenario, allows scholars to unearth pages of content written in their

language of study that would otherwise go undiscovered —or at least would be much more difficult to find—as a result of them being "masked" by appearing in a volume identified as being in a different language. We focused our study precisely on these cases.

Having randomly sampled 400 items where the volume level language metadata was *not* Japanese but the NLP language identification tool had classified a page as having Japanese text, we relied on human classification to determine the actual language of each page. Overall the accuracy of the NLP Japanese text was poor. Examples of pages erroneously identified as Japanese included: illustrations, blank pages with a few "noise" marks on them, handwritten texts, mathematical or musical notation, pages with a substantial portion of characters misidentified as Japanese *kanji*. We found the largest category of error to be scanned images that included Kanji characters that the NLP tool had classified as being Japanese when they were actually Chinese. In fact, out of the 400 sampled pages, only 1 example was found that was actually Japanese text. (Keep in mind that our sample set consisted purposefully as volume-page "mismatches" of language identification.) We then studied the opposite phenomenon, sampling pages identified as anything *other than* Japanese, from volumes human-cataloged as Japanese. This second study also surfaced a substantial number of algorithmically-introduced errors, assignable to a different set of error categories.

What is the research cost of these errors in terms of misidentified language materials? Figure 1 summarizes our initial calculations. HathiTrust contains 559,718 volumes human-identified as Japanese, consisting of 249,252,918 pages. There are also 176,300,305 pages algorithmically-identified as Japanese, spanning 623,623 volumes. The intersection of these sets is the degree of agreement between these two methods of identifying Japanese language materials: there are 168,026,395 pages in common, coming from 501,150 volumes. This mismatch indicates that scholars are likely to miss a substantial amount of text from either search methodology.

Query Term	Volumes	Pages
volumelanguage_htrcstrings:jpn	559,718	249,252,918
ja_htrctokentext:*	623,623	176,300,305
ja_htrctokentext:* AND volumelanguage_htrcstrings:jpn	501,150	168,026,395

Figure 1.
Summary of potentially "missing" Japanese-language materials between two methods of retrieval.

The error rates found through both these analyses are high enough that we are considering changes both in the Workset Builder interface (to provide caveats for researchers upon executing page-level language searches), and in the production pipeline for the next release of the EF dataset: to employ newer and different language-detection packages (an approach that appears promising in pilot tests), and to seek access to an altogether new source of language detection: that often—but not always—is provided during the initial OCR processes, and encoded in metadata not previously available to us.

While we stand by the decisions that led us to favor human language identification at the volume level, and algorithmic language identification at the page level, we are nonetheless inspired to refine and qualify both process and presentation of this important dataset.

Bibliography

[1] **Jett, J., Capitanu, B., Kudeki, D., Cole, T., Hu, Y., Organisciak, P., Underwood, T., Dickson Koehl, E., Dubnicek, R., & Downie, J. S.** (2020). *The HathiTrust Research Center Extracted Features Dataset (2.0)*. HathiTrust Research Center. https://doi.org/10.13012/R2TE-C227

[2] **Meknavin, S., Kijsirikul, B., Chotimongkol, A., & Nuttee, C.** (1998). *Combining trigram and winnow in Thai OCR error correction. COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics.*

[3] **Ikeda, K., Hayashi, R., Nagasaki, K., & Morishima, A.** (2017). *Human-assisted OCR of Japanese books with different kinds of microtasks. iConference 2017 Proceedings Vol. 2.*

[4] **Yin, Y., Zhang, W., Hong, S., Yang, J., Xiong, J., & Gui, G.** (2019). Deep learning-aided OCR techniques for Chinese uppercase characters in the application of Internet of Things. *IEEE Access*, 7, 47043–47049.

[5] The Optimaize Language Detector. https://github.com/optimaize/language-detector.

[6] **Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D.** (2014). *The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 55–60).

Creative Flows: Artistic Inspiration in/through/with Katherine Dunham's Transnational Circulation

Bench, Harmony
bench.9@osu.edu

The Ohio State University

Elswit, Kate

Kate.Elswit@cssd.ac.uk
Royal Central School of Speech and Drama, University of London

Jimenez-Mavillard, Antonio

Antonio.Jimenez-Mavillard@cssd.ac.uk
Royal Central School of Speech and Drama, University of London

Uzor, Tia-Monique

tia-monique.uzor@cssd.ac.uk
Royal Central School of Speech and Drama, University of London

This poster comes from the AHRC-funded project *Dunham's Data: Katherine Dunham and Digital Methods for Dance Historical Inquiry*. The overarching project explores the kinds of questions and problems that make the analysis and visualization of data meaningful for dance history, through the case study of 20th century African American choreographer Katherine Dunham, who toured the globe extensively, picking up performers and gathering culturally specific movement for her repertory as she went (Bench and Elswit 2020; Bench and Elswit 2022). Drawing on data-informed research in theatre history (Varela 2021; Miller 2016) and feminist and anti-racist approaches to data (D'Ignazio and Klein 2020; Johnson 2018), our manually-curated core project datasets represent Dunham's everyday itinerary of over 5000 days spent between 1947-1960 on every continent but Antarctica, the almost 200 dancers, drummers, and singers who travelled with her, and the almost 200 interconnected elements of repertory that they performed. These are currently being expanded to 1938-63, covering the majority of Dunham's stage career, and all of her domestic and international touring.

Reflecting the 2022 conference theme "Responding to Asian Diversity," we will present digital visualizations based on data collected regarding the time Dunham spent touring in the Asia-Pacific region from 1956-1958, in particular highlighting research findings related to: 1) the sites of performances and other travel to 21 cities across Australia, New Zealand, Philippines, Singapore, Malaya, Hong Kong, Korea, and Japan, 2) the performers who toured with her and those who joined her company en route (including two Australians, a New Zealander, and eight Filipinos), and 3) the existing repertory they brought with them to perform in these locations, as well as new repertory inspired by their time in the region.

The company's international touring brought them into contact with a broad range of rhythms, gestures, and referents, which then circulated onward as Dunham toured. During this period, the company began to perform a number based on the Maori haka, which was later combined with Baby San and Planting Rice, pieces that referenced travels to Japan, Korea, and the Philippines, to form Eastern Suite. We employ spatial, network, and computational analyses as they are used in performing arts research (Bollen and Holledge 2011; Balme 2019) to better understand how Dunham's choreography materializes the influence of the many geographic places that infused her diasporic imagination, and trace the flows of performers working together over time and space as a dynamic collective, and how their embodied knowledge supports the creation and transmission of Dunham's repertory. The analyses and visualizations displayed on this poster use a combination of Python/Pandas, Matplotlib, Seaborn, Gephi, Leaflet, and NetworkX. Based on Dunham's program notes for her choreography, we geolocated every repertory work (accurate to the degree Dunham described) and assigned the map a 2D color palette in such a way that repertory associated with locations of inspiration near each other will have similar colors, which we use to represent these both on the map and off as a stacked bar chart by year. We then seek to understand the multi-directional force of inspiration by connecting timelines of Dunham's places visited and of repertory inspired by place as a bipartite graph. This is further complicated by joining three datasets to examine the correlations of Dunham's travel itinerary by means of a temporal punch card, the trajectories of each company member through the company, and the passports they each carried. Together, these analyses make traceable potential ripples of Dunham's influence in the many locations the company visited, including Dunham's long term impact in the Australian entertainment landscape (Bollen 2020), and the ways in which her presence is narrated in the development of Japanese Butoh (Michio 2019).

Because scholars generally consider Dunham's artistic and political project to be one of tracing resonances and retentions of Africanist elements in diasporic movement practices throughout the Caribbean and Americas (Clark 1994; Manning 2004; Das 2017), they have not fully accounted for Asia as a site of inspiration for her choreographic work, and how her influence may have extended throughout the area as performers joined and left the company while touring, as well as the impact of her depiction of African-diasporic practices on local audiences. This poster connects with current scholarship on the African diaspora beyond the Black Atlantic (Gilroy 1993) toward various Black internationalisms that "have never been contained within the holy trinity of Europe, Africa, and the Americas" (Patterson and Kelley 2000, 32) and offers an opportunity to illuminate Dunham as part of transnational

creative flows between the Asia-Pacific region and the Afro-Caribbean and Americas.

Bibliography

Balme, C. (2019). The Globalization of Theatre 1870–1930: The Theatrical Networks of Maurice E. Bandmann . Cambridge: Cambridge University Press.

Bench, H. and Elswit, K. (2020). Katherine Dunham's Global Method and the Embodied Politics of Dance's Everyday, Theatre Survey, 61(3): 305-30.

Bench, H. and Elswit, K. (2022). Visceral Data for Dance Histories: Katherine Dunham's People, Places, and Pieces, TDR, 66(1): 39-62.

Bollen, J. (2020). Touring Variety in the Asia Pacific Region, 1946–1975 . London: Palgrave.

Bollen, J. and Holledge, J. (2011). Hidden Dramas: Cartographic Revelations in the World of Theatre Studies, The Cartographic Journal, 48(4): 226-36.

Caplan, D. (2016). Reassessing Obscurity: The Case for Big Data in Theatre History. Theatre Journal, 68 (4): 555-573.

Clark, V. (1994). Performing the Memory of Difference in Afro-Caribbean Dance: Katherine Dunham's Choreography, 1938-87. In Fabre, G and O'Meally, R. G. (eds), History and Memory in African-American Culture. New York: Oxford University Press. 188-204.

Das, J. D. (2017). Katherine Dunham: Dance and the African Diaspora. New York: Oxford University Press.

D'Ignazio, C. and Klein, L. F. (2020). Data Feminism. Cambridge, MA: The MIT Press.

Johnson, J. M. (2018). Markup Bodies: Black [Life] Studies and Slavery [Death] Studies at the Digital Crossroads, Social Text, 36(4):57–79.

Gilroy, P. (1993). The Black Atlantic: Modernity and Double Consciousness. Cambridge, MA: Harvard University Press.

Manning, S. (2004). Modern Dance, Negro Dance: Race in Motion. Minneapolis: University of Minnesota Press.

Michio, A. (2019). From Vodou to Butoh: Hijikata Tatsumi, Katherine Dunham, and the Trans-Pacific Remaking of Blackness. In Baird, B. and Candelario, R. (eds), The Routledge Companion to Butoh Performance. New York: Routledge.

Patterson, T. R. and Kelley, R. D. G. (2000). Unfinished Migrations: Reflections on the African Diaspora and the Making of the Modern World, African Studies Review, 43(1): 11-4.

Varela, M. E. (2021). Theater as Data: Computational Journeys into Theater Research. Ann Arbor: University of Michigan Press.

"The Great Wall of China" and "Hara-kiri and Feuilleton" or how to search in a digital edition of satirical texts with a focus on the theme of "an oriental fantasy especially oriented about the orient".

Biber, Hanno
hanno.biber@oeaw.ac.at
Austrian Academy of Sciences, Austria

Introduction

Digital editions have to be evaluated and examined with respect to their practical use for certain research purposes. In the following proposal an exploration of one specific digital edition will be suggested, which shall meet two aims. First, by thematically referring to the general theme of the conference, a study of the historical clichés of how the East is viewed from the West, will be carried out. Second, an investigation of the practical use of a digital edition will be done to demonstrate the feasibility for a such special research purpose of digital literary studies. The precise question put forward considers the functions of a well-developed digital edition of satirical texts for the purpose of carefully exploring the research topic of the exemplary theme mentioned above as implied in the quotation of "an oriental fantasy especially oriented about the orient".

Digital Edition and Methodology

The edition in question is the online digital edition *AAC-Fackel* (Biber 2007) of the literary journal, published on the basis of the print version of the German original that was edited and almost entirely written by the famous Viennese satirist Karl Kraus between 1899 and 1936. The digital edition of this important historical resource of *Die Fackel* enables literary scholars to study the texts in great detail, not only because of the digital accessibility of its more than six million tokens (cf. Biber 2015), but also because the numerous digital sources to be studied are provided with additional linguistic information (cf. Fischer-Starcke 2010), so that all of their lexical entities, words and phrases to be

studied, are searchable based upon technologies offered by corpus linguistics (cf. Stefanowitsch 2020).

Texts and Examples

In several cases the satirical texts by the language and social critic address the theme of the highly questionable cliché-ridden relations between orient and occident, in particular as the East is viewed from the West in newspaper articles of the time, of which many different feuilletons, essays, editorials and the like, are quoted and satirically commentated upon by the satirist. The satires from the period before 1914 are particularly relevant and interesting in this respect, because they give an impression and at the same time offer methods as how to judge the public impact of notoriously influential cultural stereotypes propagated by newspapers. Two texts by Karl Kraus are of particular concern as starting points of this demonstration: *The Great Wall of China* from 1909 and *Hara-kiri and Feuilleton* from 1912. In the first text, the true story of the murder of a young white missionary allegedly by a Chinese waiter in New York and how this crime is reported in the press, exposes and criticises the stereotypes about China and the Chinese, the hypocrisy in the context of sexual and racial morals of society. The second text about the reports of the ritual suicide of a famous general investigates the martial platitudes about Japan and the Japanese and also satirically explores and sharply criticises the language of cultural clichés and decadent aestheticism in journalism. In both texts the specific tone of voice as well as the special choice of words carrying the moral prejudices about peoples from the Far East, in the context of sex and crime as exposed and thus exploited by the media of the time, the press, are to be investigated in detail and carried out beyond these in other texts of comparable thematic scope, by making use of the instruments of a digital edition, which means to use the search mechanisms available, to use key-word-in-context lists, extract linguistic information as concerns part-of-speech, perform word form searches and reversed word form searches, investigate collocational patterns, metaphorical usage etc.

Conclusion

The limits and constraints as well as the advantages of various word searches by making use of the digital edition are going to be demonstrated. The satires by Karl Kraus expose the hypocrisy of the newspapers by using these observable yet still very powerful clichés and the bourgeois double standards of society to be followed in newspaper feuilletons, which are quoted and analysed in the texts to be studied with linguistic precision in order to give an example of how to search successfully in a digital edition.

Bibliography

Biber, H. et al. (eds.) (2007). AAC-Austrian Academy Corpus. AAC-Fackel. Online Version: Die Fackel. Herausgeber: Karl Kraus, Wien 1899-1936. https://fackel.oeaw.ac.at

Biber, H. (2015): AAC-Fackel. Das Beispiel einer digitalen Musteredition. In Baum, C. and Stäcker, T. (eds.): Grenzen und Möglichkeiten der Digital Humanities. Sonderband 1 (2015) der Zeitschrift für digitale Geisteswissenschaften, DOI 10.17175/sb001_019, https://www.zfdg.de/sb001_019

Fischer-Starcke, B. (2010). Corpus linguistics in literary analysis: Jane Austen and her contemporaries. London: Continuum

Stefanowitsch, A. (2020). Corpus linguistics: A guide to the methodology. (Textbooks in Language Sciences 7). Berlin: Language Science Press

Computational Literary Studies Infrastructure (CLS INFRA): a project to connect people, data, tools, and methods

Birkholz, Julie
julie.birkholz@ugent.be
Universiteit Gent

Börner, Ingo
ingoboerner86@gmail.com
Universität Potsdam

Chambers, Sally
sally.chambers@ugent.be
Universiteit Gent

Charvat, Vera
veramaria.charvat@oeaw.ac.at
Österreichische Akademie der Wissenschaften

Cinková, Silvie
cinkova@ufal.mff.cuni.cz

Charles University, Prague

Dejaeghere, Tess
tess.dejaeghere@ugent.be
Universiteit Gent

Dudar, Julia
dudar@uni-trier.de
Universität Trier

Ďurčo, Matej
matej.durco@oeaw.ac.at
Österreichische Akademie der Wissenschaften

Eder, Maciej
maciej.eder@ijp.pan.pl
Institute of Polish Language (Polish Academy of Sciences)

Edmond, Jennifer
jennifer.edmond@dariah.eu
DARIAH-EU

Fileva, Evgeniia
fileva@uni-trier.de
Universität Trier

Fischer, Frank
frank.fischer@dariah.eu
Universität Potsdam

Heiden, Serge
slh@ens-lyon.fr
Ecole Normale Supérieure, Lyon

Křen, Michal
michal.kren@ff.cuni.cz
Charles University, Prague

Kunda, Bartłomiej
bartlomiej.kunda@ijp.pan.pl
Institute of Polish Language (Polish Academy of Sciences)

Mrugalski, Michał
michal.mrugalski@hu-berlin.de
Humboldt-Universität zu Berlin

Murphy, Ciara
ciara.murphy@nuigalway.ie
National University of Ireland, Galway

Odebrecht, Carolin
carolin.odebrecht@hu-berlin.de
Humboldt-Universität zu Berlin

Raciti, Marco
marco.raciti@dariah.eu
DARIAH-EU

Ros, Salvador
sros@scc.uned.es
UNED, Madrid

Schöch, Christof
schoech@uni-trier.de
Universität Trier

Šeļa, Artjoms
atrjoms.sela@ijp.pan.pl
Institute of Polish Language (Polish Academy of Sciences)

Tasovac, Toma
ttasovac@humanistika.org
Belgrade Center for Digital Humanities

Tonra, Justin
justin.tonra@nuigalway.ie
National University of Ireland, Galway

Tóth-Czifra, Erzsébet
erzsebet.toth-czifra@dariah.eu
DARIAH-EU

Trilcke, Peer
trilcke@uni-potsdam.de
Universität Potsdam

Van Dalen-Oskam, Karina
karina.van.dalen@huygens.knaw.nl
Huygens Institute

Van Rossum, Lisanne
lisanne.van.rossum@huygens.knaw.nl
Huygens Institute

Abstract

The aim of this poster is to provide an overview of the principal objectives of the CLS INFRA project, its aims and structure as well as ways to get in touch.

Introduction

Just as much as the exact sciences, research in the social sciences and the humanities relies on research infrastructures: no research could be conducted without academic libraries, cultural heritage institutions, and academic and mass-market publishers. The digital turn, however, not only reshaped the theoretical and methodological frameworks in several disciplines, but it also redefined the notion of research infrastructures (see e.g. Borgman 2010, Moulin et al. 2011, Kitchin 2021). Nowadays, at least in Digital Humanities, it is hard to conduct any cutting-edge research without access to the relevant digital resources, tools to analyze them, networks of collaborating teams and individuals, and efficient communication channels to disseminate the results. In particular, this applies to computational literary studies (CLS).

An Infrastructure for CLS

With respect to the field of computational literary studies more specifically, the digital age offers challenges and opportunities for completing research on Europe's multilingual and interconnected literary heritage. At present, the landscape of literary data is diverse and fragmented. Even though many resources are currently available in digital libraries, archives, repositories, websites or catalogues, a lack of standardisation hinders how they are constructed, accessed and the extent to which they are reusable (Ciotti 2014). The Computational Literary Studies Infrastructure (CLS INFRA) project aims to federate these resources, with the tools needed to interrogate them, and with a widened base of users, in the spirit of the FAIR and CARE principles (Wilkinson et al. 2016, Carroll 2020). The resulting improvements will benefit researchers by bridging gaps between greater- and lesser-resourced communities in computational literary studies and beyond, ultimately offering opportunities to create new research and insight into our shared and varied European cultural heritage. CLS INFRA's efforts are central to catering to these urgent infrastructural needs of a growing user community. [1]

Rather than building entirely new resources for literary studies, the project is strongly committed to exploiting and connecting the already-existing efforts and initiatives, in order to acknowledge and utilize the immense human labour that has already been undertaken. Therefore, the project builds on recently-compiled high-quality literary corpora, such as DraCor and ELTeC (Fischer et al. 2019, Burnard et al. 2021, Schöch et al. to appear), integrates existing tools for text analysis, e.g. TXM, stylo, multilingual NLP pipelines (Heiden 2010, Eder et al. 2016), and takes advantage of deep integration with two other infrastructural projects, namely the CLARIN and DARIAH ERICs.[2]Consequently, the project aims at building a coherent ecosystem to foster the technical and intellectual findability and accessibility of relevant data. The ecosystem consists of (1) resources, i.e. text collections for drama, poetry and prose in several languages, (2) tools, (3) methodological and theoretical considerations, (4) a network of CLS scholars based at different European institutions, (5) a system of short-term research stays for both early career researchers and seasoned scholars, (6) a repository for training materials, as well as (7) an efficient dissemination strategy. The structure of the project with its work packages closely follows the above components of the infrastructure.

The project is delivered by a geographically balanced, complementary transnational consortium of key local and national infrastructure providers, covering the full range of the project's defined areas for integration and innovation and aligned so as to create a common infrastructural approach for computational literary studies. In particular the deep integration of both the CLARIN and DARIAH ERICs ensure the project's long term stability and sustainability.

Conclusion

The key aim of our poster is to provide a wide range of stakeholders – researchers, librarians, infrastructure providers – with an understanding of our project and with contact points for specific issues as to motivate them to get involved.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004984.

Beyond the authors of this poster proposal, we would like to acknowledge the role of the work package leads: Maciej Eder (coordinator), Justin Tonra, Christof Schöch, Karina van Dalen-Oskam, Carolin Odebrecht, Matej Ďurčo, Peer Trilcke, Frank Fischer, Julie M. Birkholz, Marco Raciti. Find out more about the team here: https://clsinfra.io/about/ourresearchers/.

Bibliography

Borgman, Christine. 2010. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, Mass. & London: MIT Press.

Burnard, Lou, Christof Schöch, and Carolin Odebrecht. 2021. In search of comity: TEI for distant reading. *Journal of the Text Encoding Initiative*14. https://doi.org/10.4000/jtei.3500.

Ciotti, Fabio. 2014. Digital literary and cultural studies: the state of the art and perspectives. *Between*4/8, 1-17. https://doi.org/10.13125/2039-6597/1392.

Eder, Maciej, Rybicki, Jan and Kestemont, Mike. 2016. Stylometry with R: a package for computational text analysis. *R Journal*, 8(1): 107-21. https://journal.r-project.org/archive/2016/RJ-2016-007/index.html

Fischer, Frank, Ingo Börner, Matthias Göbel, Andrea Hechtl, Christopher Kittel, P. Miling, and Peer Trilcke. 2019. Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. In *Book of Abstracts of the Digital Humanities Conference 2019*. Utrecht: ADHO.

Heiden, Serge. 2010. The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In *24th Pacific Asia Conference on Language, Information and Computation* (pp. 10 p.). Sendai, Japan. Retrieved from http://halshs.archives-ouvertes.fr/docs/00/54/97/64/PDF/paclic24_sheiden.pdf

Kitchin, Rob. 2021. *The data revolution: big data, open data, data infrastructures and their consequences*. 2nd edition. Thousand Oaks: Sage Publications Ltd.

Moulin, Claudine, Arianna Ciula, and Julianne Nyhan. 2011. *Research Infrastructures in the Digital Humanities*. Science Policy Briefing 42. Strasbourg: European Science Foundation.

Schöch, Christof, Tomaz Erjavec, Roxana Patras, and Diana Santos (to appear). Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives. *Modern Languages Open*.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3(1). https://doi.org/10.1038/sdata.2016.18.

Notes

1. Project website: https://www.clsinfra.io.
2. See https://www.dariah.eu and https://www.clarin.eu.

What's new about HuNI?

Burrows, Toby

toby.burrows@uwa.edu.au
University of Oxford, United Kingdom; University of Western Australia

Verhoeven, Deb

deb.verhoeven@ualberta.ca
University of Alberta

Over the past ten years, humanities researchers in Australia have established a large interconnected database of cultural information and used it to create detailed networks of relationships between the people, places and objects it describes. This aggregated knowledge base is called the Humanities Network Infrastructure (HuNI — "honey"): https://huni.net.au/ There are nearly eighteen million nodes in the HuNI knowledge graph, drawn from thirty-three data sources which reflect the perspectives of different disciplines across the humanities and creative arts with new data sources being added each year. HuNI uses the neo4j graph database software to create and manage its knowledge graph.

Using HuNI is not just a matter of exploring the knowledge graph through filtering and browsing. Users can actively create links between nodes, assemble nodes into personal collections, and download these collections for use in other environments. HuNI collections are a form of categorization or classification which feeds back into the graph itself. There is no prescribed ontology or taxonomy for categorizing entities; users can invent their own approach and share it with others. Among the 180 public collections constructed so far are some based around specific people ("James Cassius Williamson 1870-1913: networks and connections"), some based on specific places ("Efate", in the New Hebrides), and some based on themes and topics: "Blast Si(gh)te: bodies in time-space" – a collection of entities related to the effect of nuclear testing on indigenous people in Outback South Australia in the 1950s – and "Australian fashion 1850-1950".

HuNI has been designed to support the underlying precepts of humanities research, especially complexity, contestation, and connection. Multiple interpretations of relationships – including conflicting ones – can be represented. Connections do not have to be "right" or "authoritative", or even logical. User-generated links can be as creative and complex and nuanced and contrary as the multitude of HuNI users themselves. HuNI also encourages other users to contest these links and provide alternative interpretations of how entities are connected.

This poster presents and demonstrates several recent important developments which extend and transform HuNI in significant new ways. A CSV upload feature enables HuNI users to contribute new nodes to the HuNI knowledge graph. Previously, users have only been able to create links between existing nodes, but now they will able to add collections of nodes as well. This will significantly enhance the ability of users to shape the HuNI knowledge graph and to express their own perspectives in it.

From 2022, Canadian data will begin to appear alongside the existing Australian data. This will enable links and connections to be made between the knowledge of two countries which share a common British colonial past (Donaghy 1995), and in which the indigenous knowledge of Canadian First Nations and Australian Aboriginal and Torres Strait Islander peoples is of comparable importance (Smith 2021).

Programmes at several universities including University of Melbourne and University of Technology Sydney (UTS) have recently been developed for using HuNI in teaching relational approaches in digital historical research. At the University of Alberta, HuNI is used within courses offered to students of Digital Humanities, Women's and Gender Studies, and Library and Information Science to understand how data and power are mutually implicated, especially when data is integrated, exchanged and interoperated (Posner 2015). Using HuNI, students experiment with vernacular ontologies to explore how the relational capacity of data might also play a role in social change.

There are very few experimental humanities tools that have survived for ten years, as HuNI has. This is especially significant since the more that users contribute to the HuNI knowledge graph the greater the value of the application. HuNI enables researchers to both browse and create knowledge — through a combination of human- and system-generated "connections". To this vision HuNI also contributes the additive nature of graph databases in which new kinds of relationships can be proposed and included without disrupting the overall functionality of the knowledge base. The HuNI knowledge graph is designed for the kind of information "meandering" and serendipitous encounters which are at the heart of humanities research (Verhoeven 2016). The latest enhancements reinforce these design goals, while greatly extending HuNI's reach into the international user community.

Bibliography

Donaghy, G. (1995). *Parallel Paths: Canada-Australian Relations since the 1890s*. Ottawa: Historical Section, Department of Foreign Affairs and International Trade.

Posner, M. (2015). The radical potential of the Digital Humanities: the most challenging computing problem is the interrogation of power. *The Impact Blog, The London School of Economics and Political Science*, http://blogs.lse.ac.uk/impactofsocialsciences/2015/08/12/the-radical-unrealizedpotential-of-digital-humanities/ (accessed 16 April 2022).

Smith, L. T. (2021). *Decolonizing Methodologies: Research and Indigenous Peoples*. 3rd ed. London: Bloomsbury.

Verhoeven, D. (2016). As luck would have it: serendipity and solace in digital research infrastructure. *Feminist Media Histories,* 2(1): 7–28.

"Es war einmal …" – First Sentences in Literature: A German-Language Reference Corpus

Busch, Anna
annabusch@uni-potsdam.de
Theodor-Fontane-Archiv, Universität Potsdam

Roeder, Torsten
dh@torstenroeder.de
Bergische Universität Wuppertal, Germany

Idea

In literary and linguistic studies, the first sentence of a narratological context is a regularly studied object (on this, among others, Alt 2020, Haubrichs 1995, Hirdt 1974, Queng 2019, Miller 1965, Neuhaus 2019, Raulff 2019, Retsch 2000, Selbmann 2019). This is hardly surprising, since the first sentence has been regarded since Wolfgang Iser's study The Act of Reading as the entrance into the text through reading, as the key point of interaction between text and reader (1976: 38). In the richness of its various forms, the first sentence reveals "the treasures of literature in nuce" (Alt 2020: 18) and, with Alain Robbe-Grillet, it could be put forward that literary history is to be written from the study of its opening sentences (1992: 38).

A systematic, digitally supported study of "first sentences" has yet to be carried out. Occasionally, corpora of first sentences in German have been collected by hand (Beck 1992, Beck 1993, Wolkersdorf 1994) and attempts have been made to draw up a typology of the first sentence in literature on the basis of selected individual analyses

(most recently Alt 2020). A systematic categorisation on the basis of a semi-automated, larger corpus of research – as presented here – seems helpful. There are similar studies that inquire into the quintessence of the poetic in literature through its countability (cf. for example Moretti 2009, also Fischer/Strötgen 2015, Fischer/Jäschke 2018a/b); a single quantifying study dealing decidedly with German-language narrative beginnings (not first sentences) can be found in the work of Herrmann 2018.

The aim of the corpus "First Sentences in German-Language Literature" is to address the "lack of an overall view" (Alt 2020: 246) of all previous studies on first sentences. To this end, a data corpus is created and published, on which an initial evaluation will be undertaken in an interlocking of quantitative and text-analytical approaches.

Project

Several full-text, open access corpora (*Deutsches Textarchiv*, *Zeno*, etc.), from which texts were extracted according to genre, serve as source material. It is clear that although the existing full-text offerings provide varying degrees of structural information about the respective document, the automatic delimitation of closed text units is often non-trivial and not possible reliably without individual examination (e.g. in the case of anthologies, texts with several chapters, texts in several volumes). However, this is the prerequisite for extracting the first sentences. In addition, the beginning of the "poetic text" cannot always be clearly localised automatically, e.g. due to prefaces, dedication texts or introductions.

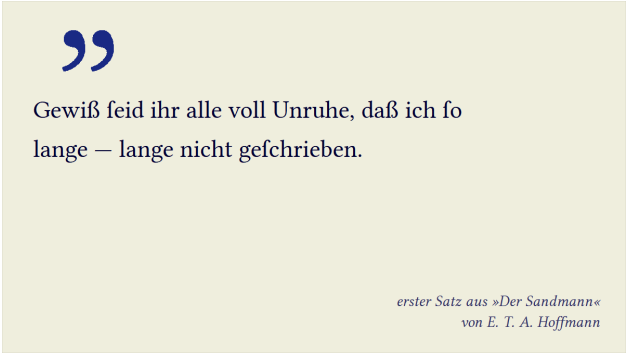
Furthermore, the delimitation of "first sentences" is a semantic problem. Sentences can be understood as grammatical-analytical units that are separated from each other by certain punctuation marks, which accommodates machine processing. However, the signs used to delimit a sentence differ and change considerably. The absolute selectivity of some punctuation marks is also questionable depending on the context, which is why sentences are sometimes to be understood as units of meaning in which punctuation marks have a structuring but not interrupting function (cf. fig. 2a/b). Should we therefore rather speak of a flowing "beginning" or "start"? Thus, areas of vagueness play into the determination of "first sentences", which in turn can affect corpus consistency and comparability.

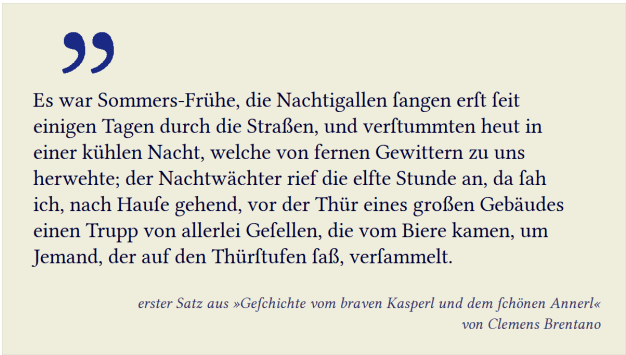
Evaluation

The currently created corpuses of novels, novellas and fairytales is completely encoded in TEI, including metadata and source information, including positional information (available at https://github.com/satzomat/corpus). Depending on the genre, the number of first sentences ranges between 100 and 1,000 entries. With the help of the manually and automatically created annotations, the corpus can be analysed and visualised according to various parameters, such as date of publication, text genre, gender of author, references to persons, places or time in the text (cf. Fig. 1c) or length of the entire text. In addition, it is documented which selection criteria the respective data sources were subject to and how this should be taken into account in the evaluation with regard to the balance of the corpus (cf. Hug/Boenig 2021). To disseminate the corpus, the Twitter project @satzomat was launched in 2021, which sends two first sentences daily (cf. Figures 1–3).

The aim is to create a "typology of incipits" with the help of computer-philological evaluation methods and to ask to what extent genres determined certain types of first sentences in the course of history (e.g. landscape image, frame story) and whether further correlations can be determined with the help of the metadata and annotations (see the project page http://satzomat.de/ for more information).

Figures





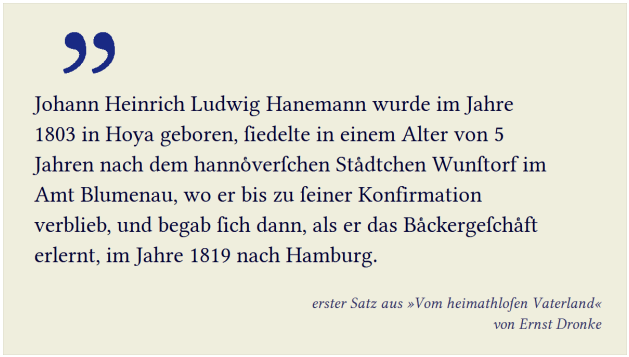
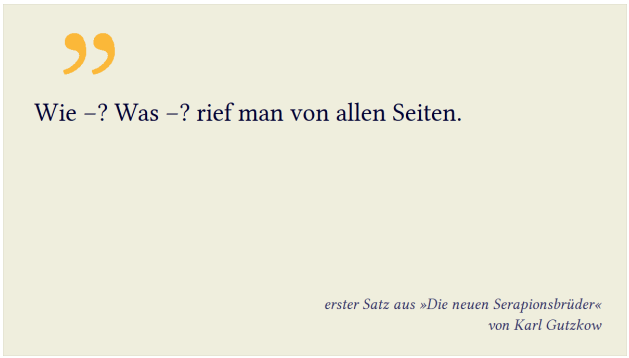
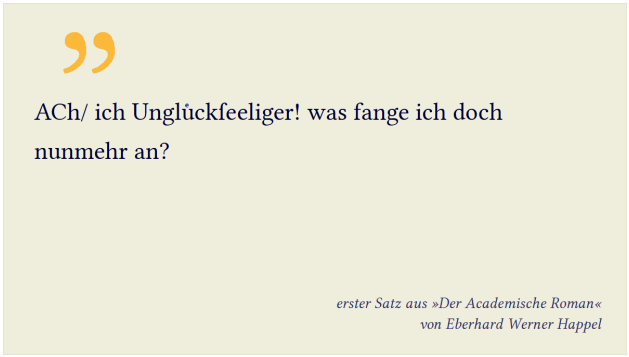


Fig. 1 a/b/c:
Novella beginnings





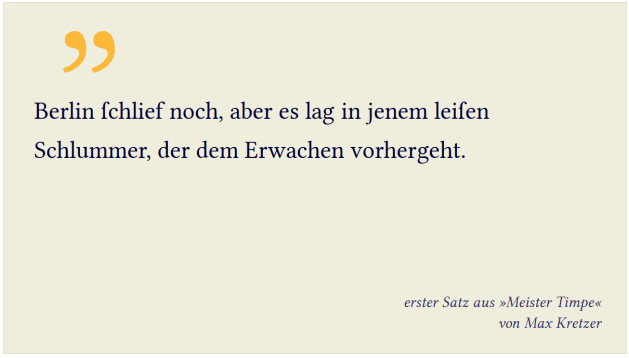
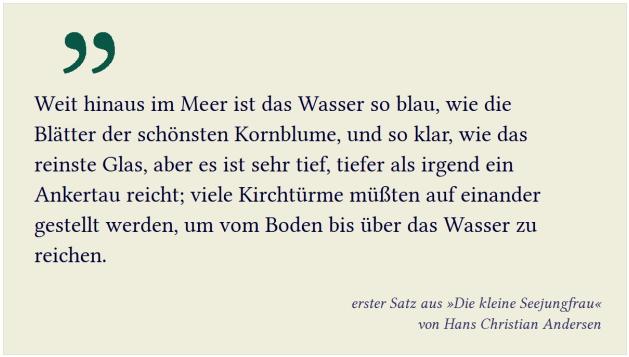
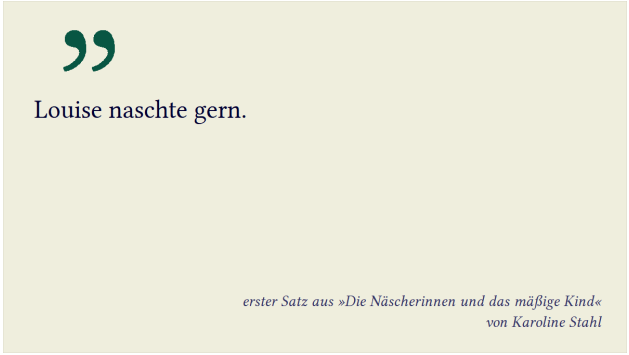


Fig. 2 a/b/c:
Novel beginnings





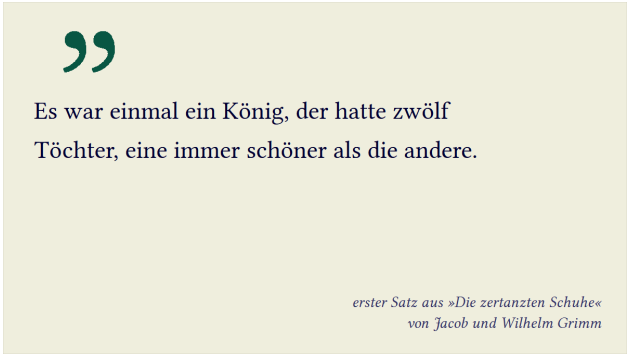


Fig. 3 a/b/c:
Fairy tale beginnings

Bibliography

Alt, Peter-André (2020): *'Jemand musste Josef K. verleumdet haben …' Erste Sätze der Weltliteratur und was sie uns verraten.* München: Beck.

Beck, Harald (1992)**:** *Roman-Anfänge. Rund 500 erste Sätze*. Zürich: Haffmans.

Beck, Harald (1993)**:** *Romanenden. Rund 500 letzte Sätze*. Zürich: Haffmans.

Fischer, Frank / Strötgen, Jannik (2015): "Wann findet die deutsche Literatur statt? – Zur Untersuchung von Zeitausdrücken in großen Korpora." Presented at the DHd2015 *Von Daten zu Erkenntnissen: Digitale Geisteswissenschaften als Mittler zwischen Information und Interpretation.* 2. Tagung des Verbands "Digital Humanities

im deutschsprachigen Raum" (DHd2015), Graz: Zenodo. http://doi.org/10.5281/zenodo.4623384 [last access: 9. December 2021]

Fischer, Frank / Jäschke, Robert (2018a): "Liebe und Tod in der Deutschen Nationalbibliothek. Der DNB-Katalog als Forschungsobjekt der digitalen Literaturwissenschaft." Presented at the DHd 2018 Kritik der digitalen Vernunft. 5. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum" (DHd 2018), Köln: Zenodo. http://doi.org/10.5281/zenodo.4622376 [last access: 9. December 2021]

Fischer, Frank / Jäschke, Robert (2018b): "Ein Quantum Literatur. Empirische Daten zu einer Theorie des literarischen Textumfangs." DFG-Symposium "Digitale Literaturwissenschaft". Villa Vigoni, 9.–13. Oktober 2017. [unpublished]

Haubrichs, Wolfgang (1995): "Kleine Bibliographie zu "Anfang" und "Ende" in narrativen Texten (seit 1965)", in: Zeitschrift für Literaturwissenschaft und Linguistik 25, 99: 36-50.

Herrmann, Berenike (2018): "Anschaulichkeit messen. Eine quantitative Metaphernanalyse an deutschsprachigen Erzählanfängen zwischen 1880 und 1926", in: Köppe, Tilmann / Singer, Rüdiger (eds.): *Show, don't tell: Konzepte und Strategien anschaulichen Erzählens.* Bielefeld: Aisthesis 167-212.

Hirdt, Willi (1974): "Incipit. Zu einer Poetik des Romananfangs", in: Romanische Forschungen LXXXVI: 419-436.

Hug, Marius / Boenig, Matthias (2021): Die Geschichte der Digitalen Bibliothek, oder: Aller guten Kurationen sind drei+: https://sprache.hypotheses.org/2436 [last access: 9. December 2021]

Iser, Wolfgang (1976): Der Akt des Lesens. Theorie ästhetischer Wirkung. München: Fink.

Miller, Norbert (1965): *Romananfänge. Versuch zu einer Poetik des Romans.* Berlin: Verl. Literarisches Colloquium.

Moretti, Franco (2009): Style, Inc Reflections on Seven Thousand Titles (British Novels, 1740-1850), in: *Critical Inquiry* 36, I: 134-158.

Neuhaus, Stefan (2019): "Aber wehe, wehe, wehe! Wenn ich auf das Ende sehe!!" Wie in Romanen und Erzählungen durch Anfang und Ende ein Rahmen erzeugt wird, in: Neuhaus, Stefan / Weber, Petra (eds.): *Anfangen und Aufhören*. Paderborn: Wilhelm Fink 141-157.

Queng, Jesse (2019): "Syntaktische Strukturen als poetologisches Mittel des Anfangens in der Prosa: Der erste Satz von Heinrich Bölls Irischem Tagebuch", in: Neuhaus, Stefan / Weber, Petra (eds.): *Anfangen und Aufhören*. Paderborn: Wilhelm Fink 89-101.

Raulff, Ulrich (2019): "Letzte Sätze", in: *Zeitschrift für Ideengeschichte* 13: 129-142.

Retsch, Annette (2000): *Paratext und Textanfang.* Würzburg: Königshausen & Neumann.

Richardson, Brian (2008): *Narrative Beginnings: Theories and Practices*. University of Nebraska Press.

Robbe-Grillet, Alain (1992): "Warum und für wen schreibe ich", in: Bühler, Karl Alfred (ed.): *Robbe-Grillet zwischen Moderne und Postmoderne - "nouveau roman", "nouveau cinéma" und "nouvelle autobiographie*". Tübingen: Narr.

Selbmann, Rolf (2019): "Lauter erste Sätze", in: Neuhaus, Stefan / Weber, Petra (eds.): *Anfangen und Aufhören*. Paderborn: Wilhelm Fink 67-87.

Wolkersdorfer, Andreas (1994): *Der erste Satz. Österreichische Romananfänge 1960-1980*. Wien: WUV Univ.-Verl.

MIV17: a database for 17th-century manuscript culture

Crespi, Serena Carlamaria

serenacarlamaria.crespi@gmail.com

Our knowledge of 17th century Italian manuscript culture is very limited, given the complexity of the surviving documents, with most textbooks and contributions ending their investigation towards the end of the 16th century. However, thanks to the employment of digital tools it is now possible to reconstruct the history of culture and manuscript circulation of this century.

This proposal aims at showing the current progress of my doctoral research, based on the reconstruction of the seventeenth century Florentine manuscript culture. During my research I have collected and implemented 4.814 bibliographic records of Italian manuscripts that are preserved in the most important Florentine libraries (BNCF, Laurenziana, Riccardiana, Moreniana and Marucelliana). These data, of both a codicological and textual nature, have been used for statistical and quantitative analysis with the aim of reconstructing the Florentine manuscript culture of the 17th century, thanks to the new methods of distant reading.

The project started by cataloguing more than four thousand 17th century manuscripts that were kept in several libraries in Florence [fig. 1].

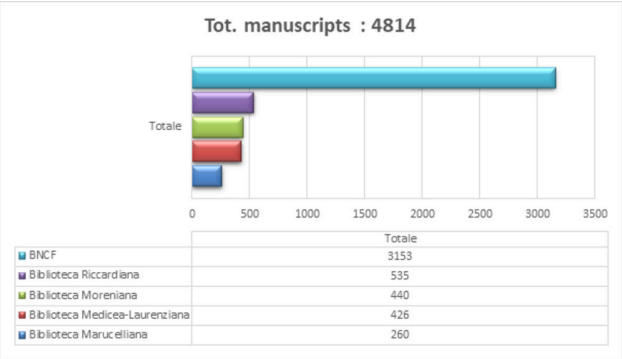


Fig. 1
A sample of the manuscript corpus in each library

Following the idea that "a large number of manuscripts cannot be studied in the same way as a single one" (Derolez, 2006), the initial corpus of manuscripts was parceled out into an Excel set of records and then was transformed into a database (the so-called (*MIV17*), using the XML-TEI schema. The choice of using the XML-TEI standard was made to simplify the future sharing of data with other similar digital projects and archives, such as MOL (*Manus OnLine)*, while ensuring the durability of the data. The TEI model for MS Description was then modified and the dataset was implemented using the EAC-CFP standard to create and connect Authority and Corporate Bodies files, containing any information about manuscript responsibility and Florentines cultural environment.

This large corpus of bibliographic and codicological information was used as the starting environment for statistical and quantitative analysis. By exploring the possibilities given by the distant reading methods it was possible to match and picture the recurrence of common characteristics, like the relation between the support dimension and the literary genres, the typologies of texts subjected to the handwritten circulation, the mix between printed and handwritten parts in the same *codex*, the presence of dedications and their relation with the format or the type of book.

On the other hand, a more socio-cultural focused approach led me to define two possibilities of circulation: the so-called horizontal one (for copies and transmissions between academics, literary congregations, acquaintances), and the vertical one, which intends the manuscript as an object of luxury, collection or a gift.

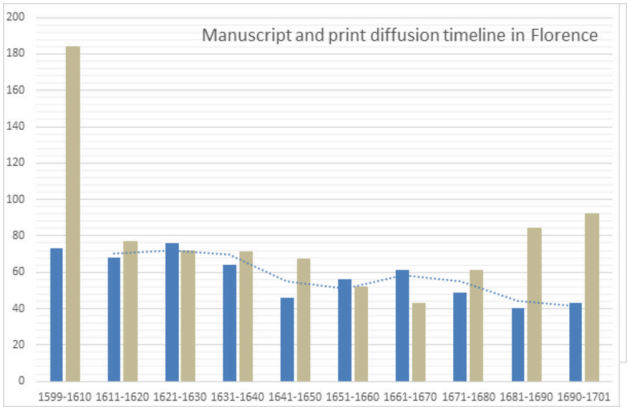


Fig. 2
Range of manuscript (blue) and print (grey) circulation

These preliminary studies have also led to highlight certain points of interest, such as the methods of handwritten transmission in a specific cultural environment, like Literary Academies, or the existence of social networks of readers, and they were useful to underline few aspects of the complex relation between the printed book and the manuscripts [see fig. 2].

Although these studies are still in the experimental phase, sharing the collected data can bring new information and stimulate new approaches to the study of the manuscript culture in the 17th-century Italy. Divulgation is, in fact, another central point of this project; an open access interface will be published to give researchers the possibility to use the collected data, combining targeted searches and statistical matches, following the idea of micro and macro analysis.

Bibliography

Burlinson, Christopher, «Manuscript and Print, 1500-1700», Oxford Handbooks, Online (2016): DOI:10.1093/oxfordhb/9780199935338.013.86.

Derolez, Albert, «The Codicology of Italian Renaissance Manuscripts: Twenty Years After». pp. 233-240, Manuscripta 50.2 (2006).

Jocker, Matthew, L. «Macroanalysis: Digital Methods & Leteraty History». University of Illinois Press, (2013).

McKitterick, David. «The invention of Rare Books: Private Interest and Public Memory, 1600-1840». Cambridge University Press (2020).

McKitterick, David. «Print, Manuscript and the search of order, 1430-1830». Cambridge University Press (2006).

Moretti, Franco. «Distant reading». Verso books (2013).

Richardson, Brian, «Manuscript Culture in Renaissance Italy» Cambridge University Press, (2009).

Stephan, Jänicke, Greta, Franzini, Muhammad Faisal, Cheema e Gerik Scheuermann «On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges». Eurographics Conference on Visualization (EuroVis) (2015), R. Borgo, F. Ganovelli, and I. Viola (Editors) (2001): https:/ /www.informatik.uni -leipzig.de/ ~stjaenicke/Survey.pdf

Establishing a Code Review Community for DH

Damerow, Julia
jdamerow@asu.edu
Arizona State University

Sutton Koeser, Rebecca
rebecca.s.koeser@princeton.edu
Princeton University

Gao, Andrew
andrewgao22@gmail.com
Canyon Crest Academy

Vogl, Malte
mvogl@mpiwg-berlin.mpg.de
Mac Planck Institute for the History of Science

Zandbank, Itay
itay@researchsoftware.co.il
The Research Software Company

Tharsen, Jeffrey
tharsen@uchicago.edu
The University of Chicago

Casties, Robert
casties@mpiwg-berlin.mpg.de
Mac Planck Institute for the History of Science

Westerling, Kalle
kalle.westerling@gmail.com
British Library

Carver, Jeffrey
carver@cs.ua.edu
University of Alabama

Many digital humanities projects require custom software development. The people doing that work, who may be researchers, professional software developers, or someone in between, write software to achieve a project's goals. But who is testing and reviewing that software to confirm it works properly? No matter how experienced or well-trained a programmer is, there will inevitably be errors in the produced code. A rough estimate suggests that there are between 15 to 50 errors in 1000 lines of code written by professional software developers (Soergel 2015). While not all errors affect the research findings based on the code, it is possible, and there are plenty of cases where this has happened (see for example Letzner et al. 2020 or Miller 2006). Furthermore, uncaught "edge cases" could drastically affect future researchers' results.

Code review is a widespread technique to improve software and reduce the number of flaws. In a code review, a programmer (other than the original code author(s)) reviews the source code. They ask questions and make suggestions for improving the software. In addition to identifying and eliminating errors, code review can improve overall quality by making the source code more readable and maintainable. Furthermore, code reviews can improve not just the skills of the reviewee but also those of the reviewer. If a code author and reviewer work in the same team or on the same or related projects, code reviews can also support team cohesion and facilitate information-sharing within a team.

Code reviews are fairly easy to implement in teams of two or more developers, where there is a shared context, technical stack, and agreed upon conventions. However, in digital humanities projects, often there is just the one "techy" person who does all the coding with no colleague to review their code. Given the prevalence of virtual communication platforms like Slack and Github, there is no reason that code review may only happen internally at a single lab. Rather, programmers across labs/centers may review each other's code. At the ACH 2021 conference, a group of people organized a workshop to discuss and develop ideas and strategies for a community code review process for digital humanities. The outcome of the workshop was a working group as part of the ADHO SIG DHTech that meets monthly with the goal of building a community of people and a community code review infrastructure.

A community code review system would provide developers writing code for digital humanities projects with a way to ensure the quality of their code. Similarly, it would give researchers or developers reusing code reviewed software some insurance that a program generates trustworthy results. It would also fill a gap in the current publishing landscape that consists of journals like the Journal of Open Source Software (https://joss.theoj.org/)

that provides ways for developers to publish about the software they create. These journals typically require software to be "feature-complete" and ideally reusable. They check that certain best practices are followed (such as providing installation instructions or API documentation) but for good reason a full-fledged code review is usually not possible. Additionally, a community code review system would provide graduate students doing computational research in their dissertation but who do not have a technical reader on their committee, or researchers who begin using computational methods with a way to get feedback on their programming work. This proposal is for a virtual poster that describes the work of the working group and its goals.

Bibliography

Letzner, S., Güntürkün, O., and Beste, C. (2020). Retraction Notice to: How Birds Outperform Humans in Multi-Component Behavior. *Current Biology*. https://www.cell.com/current-biology/comments/S0960-9822(17)30960-0. Accessed December 8. 2021.

Miller, G. (2006). A Scientist's Nightmare: Software Problem Leads to Five Retractions. *Science* 314 (5807): 1856–57.

Soergel, D. A. W. (2015). Rampant Software Errors May Undermine Scientific Results. *F1000Research* 3 (303). https://doi.org/10.12688/f1000research.5930.1.

DHTech - An ADHO Special Interest Group

Damerow, Julia
jdamerow@asu.edu
Arizona State University

Vogl, Malte
mvogl@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science

Casties, Robert
casties@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science

Gao, Andrew
andrewgao22@gmail.com
Canyon Crest Academy

Sutton Koeser, Rebecca
rkoeser@princeton.edu
Princeton University

Tharsen, Jeffrey
tharsen@uchicago.edu
The University of Chicago

Zandbank, Itay
itay@researchsoftware.co.il
The Research Software Company

In 2017, a group of software developers, scholars with programming expertise, and project managers got together at a workshop at the DH 2017 conference in Montreal. The topic of the workshop was the development of an infrastructure for collaboration and cooperation in regard to tool development. It resulted in the establishment of DHTech (https://dh-tech.github.io/), a community for people doing technical work in DH to exchange knowledge, share expertise, and foster collaboration among Digital Humanities software projects.

Fast forward four years, in early 2021, DHTech became an ADHO Special Interest Group (SIG). DHTech has a steering committee currently consisting of seven people that organizes activities related to DHTech. It also has working groups that give members of the community the opportunity to collaborate on specific topics of interest. Furthermore, DHTech organizes virtual meetups and workshops to discuss technical and other topics the members of DHTech are interested in.

We are proposing a virtual poster at DH2022 that introduces DHTech to the wider DH community. Currently, DHTech has around **140 members** and is growing slowly but consistently. We strongly believe that joining the DHTech community is beneficial for anyone doing technical work in the Digital Humanities as it offers a way for exchanging ideas and knowledge, getting support, and connecting with like-minded colleagues. In the Digital Humanities, where there is often just one person in a lab, department, or project responsible for the technical part of a project, a community like DHTech can be highly valuable by providing a sounding board and offering input and ideas. DHTech community members are eager to support each other.

In the last four years, DHTech has, among other things, successfully accomplished the following tasks:

- Held several virtual workshops and meet-ups discussing not just technical topics but also issues like onboarding of new DH developers or building local DH communities (recordings can be found here).
- Maintained a website with blogs and news entries as well as a Slack workspace that allows its members to

share experiences, accomplishments, and events, and to connect with each other.
- Conducted a survey regarding technical skills and demographics of people doing technical work in DH. The results of the survey can be found here.
- Established an active working group that aims to ensure code quality of DH projects by developing a peer code review system.
- Organized several workshops at conferences to discuss topics of interest. For example, one resulted in a white paper to clarify the role of DH developers in humanities research, which can be found on the DHTech website. Another workshop on code reviews resulted in the above-mentioned working group.

Future plans for DHTech include a second working group that will focus on a technical mentorship program for DH developers, yearly steering committee elections to ensure long-term sustainability of the community, and the establishment of a peer code review infrastructure. We hope that a virtual poster at DH2022 will help to grow the DHTech community even further and to increase the diversity of its members.

Fiction, Data: Distant Reading of the Hebrew Novel

Dekel, Yael
yaelde@bgu.ac.il
Ben Gurion University, Israel

This proposed poster is dedicated to my ongoing project "*Roman Mafte'ach:* Distant Reading in the Hebrew Novel", which constitutes applied research in Hebrew literature and digital humanities, ultimately aiming to create a comprehensive, up-to-date database of the Hebrew novel. My project approaches the Hebrew novel from a bird's eye view, along the lines of the literary-historical approach advanced by Franco Moretti as "distant reading" (Moretti, 2000), balanced carefully with more traditional hermeneutic approaches, as conceptually described by Jan Christoph Meister (2014).

The scope of this project is as broad as the scope of the Hebrew novel, as I collect data on every novel originally published in Hebrew, venturing out beyond the canon, and attempting to cover "the great unread" (Margaret Cohen, 1999). *Roman Mafte'ach*, therefore, extends from Avraham Mapu's *Ahavat Zion* (1853) which is considered the first Hebrew novel, to the present, and includes the latest Hebrew novels published today, in Israel and elsewhere.

As Marienberg-Milikowsky has pointed out, "From a purely quantitative standpoint, Hebrew (prose) literature consists of a relatively small corpus." This offers an advantage, allowing for the back-and-forth movement between distant and close reading, between personal literary interpretation and collective, quantitative analysis (Marienberg-Milikowsky, 2019).

The backbone of *Roman Mafte'ach* is a platform for providing computer-aided analysis of the Hebrew novel, based on two complementary yet independent steps (both will be presented in the poster):

1. An inventory of all the titles and authors of the Hebrew novel, from *Ahavat Zion* to the present day. With the support and assistance of staff from the National Library, I compiled an initial inventory of the Hebrew novel, carefully sifting through it until it now consists of roughly 8,500 titles. It is important to stress that the list includes data that does not exist elsewhere. This is because, up until my research, the Hebrew novel was not seen as a distinct category in library catalogues and databases. The importance of this list is twofold: first, the inventory itself is a source of data (mainly of bibliographical nature: titles, authors, year of publication, publishing house, number of pages). Thus, it answers several of the main questions that I posed when initiating the project – the most crucial one being: How many titles make up the sphere of the Hebrew novel? Second, the list is valuable for validating, as well as anchoring, some of the data I collect using questionnaires.
2. Individual responses to questionnaires I deliver to different readers, both professional and non-professional. The questionnaire is designed to collect data in several main categories, using multiple-choice questions, linear scales, and a few short-answer questions which allow for more personal and interpretive responses. The categories of the questionnaire are: **Bibliography** (name of author, title, year of publication, publishing house, editor, number of pages);**Structure** (Sub-genre; graphic components; sub-structure of the novel; chapter length);**Narratological aspects** (type of narrator, key-events, pacing of narrative, types of exposition and closure); **Language** (grammatical tense, linguistic register, other languages used in the novel, inter-textuality);**Time and space** (temporal scope of the novel, main historical epoch, main space, main geographical area); **Themes** (this part includes multiple choice questions about many themes addressed by the Hebrew novel, e.g., love, family, childhood, marriage, physical illness, mental illness, crime, Judaism, Christianity, religiosity, war, sex, science and technology, climate change and more).

This is a special form of distant reading that I term "public reading" (Dekel and Marienberg-Milikowsky, 2021).

Such a project – in scale, methodology and aims – has never before been carried out in the field of Modern Hebrew literature nor, to the best of my knowledge, in other similar fields or in other languages-literatures. This is a historiographical project; therefore, it builds on existing historiographies of Hebrew literature (*inter alia* by Shaked, Miron and Schwartz) and yet it aims to show the picture in a different way. In the poster, I will present – using graphs that include some of the respective data – the two components of the project. Moreover, I will also reflect on the complexities that such a method provokes.

Bibliography

Margaret Cohen (1999): *The Sentimental Education of the Novel*, Princeton University Press.

Yael Dekel and Itay Marienberg-Milikowsky (2021): "From Distant to Public Reading: The (Hebrew) Novel in the Eyes of Many", *magazén | International Journal for Digital and Public Humanities*(forthcoming).

Itay Marienberg-Milikowsky (2019): "Beyond digitization? Digital humanities and the case of Hebrew literature," *Digital Scholarship in the Humanities*, Oxford Universiy Press, 2019, 908-913.

Jan Christoph Meister (2014): "Toward a Computational Narratology." In: Agosti Maristella and Tomasi Francesca. *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*: *CLEUP* , pp. 17–36.

Franco Moretti (2000): "Conjectures on World Literature", New Left Review 1, 54-68.

Replicating The Riddle of Literary Quality: The litRiddle package for R

Eder, Maciej
maciej.eder@ijp.pan.pl
Institute of Polish Language (Polish Academy of Sciences)

Lensink, Saskia
s.e.lensink@gmail.com
Independent

Van Zundert, Joris
Joris.van.zundert@huygens.knaw.nl
Huygens ING - KNAW, Netherlands, The

Van Dalen-Oskam, Karina
karina.van.dalen@huygens.knaw.nl
Huygens ING - KNAW, Netherlands, The

Our poster introduces an R package with a varied set of data and a companion website to enable replication of data analyses and findings from the project The Riddle of Literary Quality (2012-2019). The project searched for linguistic and stylistic patterns in modern Dutch novels and novels recently translated into Dutch that may help explain the literary value readers do or do not attribute to these novels. Are novels that readers consider to be highly literary measurably different from those that receive low scores for literary quality?

To answer this question, the project team built a corpus of 401 novels published for the first time in Dutch from 2007 to 2012 and most sold or most borrowed from public libraries in the period 2010-2012. Readers' opinions about these novels were gathered in 2013 in The National Reader Survey that drew 13,784 respondents. Correlating measurements of the digital texts with the opinions of the readers resulted in important knowledge about the perceptions of literariness in the Netherlands in 2013.

The survey showed how readers are biased in several ways in their attribution of literary value to contemporary novels. Books presented as genre fiction are implicitly denied their chances of literary fame. Novels labelled as literary fiction by the publisher have a far better chance of being regarded as having high literary quality. But also in this category biases apply. Female authors, for instance, have less prestige than their male colleagues, even when the linguistic features of their writing do not significantly differ. The results also showed that the level of linguistic and topic complexity correlates strongly with literary quality scores. The higher the scores for literary quality, the more difficult the books are. The project yielded two PhD-theses in English, by Andreas van Cranenburgh (2016) and Corina Koolen (2018).

We created an R package named "litRiddle" that is available through CRAN. It contains four data tables with a number of functions to access them:

- The table Books contains the names of the authors and the titles of the 401 novels in the corpus, including metadata indicating the genre label, if the book is translated or not, the gender of the author, the country of origin of the author, the original language of the book, and more. The table includes a set of linguistic measurements for each of the books.

- The table Frequencies lists the word frequencies of the 5,000 most frequent words across 401 novels. Due to copyright restrictions the litRiddle R-package cannot include the full texts of the novels. It provides relative word frequencies lists per novel that allow to replicate many of the measurements done by the research team and that will help to explore other perspectives based on bag-of-words approaches.
- The table Respondents consists of metadata for the 13,784 respondents who took part in The National Reader Survey and includes information about, for example, their age, gender, and level of education. It also includes their answer to a number of questions going into their reading habits and how they value reading.
- The table Reviews captures how respondents scored the books they had read and how they scored books that they had not read but had an opinion about. Respondents scored books for their general quality (i.e. how good the respondent judge the book to be) and their literary quality (i.e. how literary the respondent thought the book was). The table also includes the optional motivation for one of their scores if the respondent provided one. The motivations are added for 12,367 out of the total number of 448,055 individual reviews; their length ranges from one word (e.g. 'no') to one paragraph.

The litRiddle package includes information on how to use the data, combine the different tables, and produce graphs from query results.

Project leader Karina van Dalen-Oskam published a Dutch-language synthesis of the results from The Riddle of Literary Quality in 2021 and is currently working on a shorter English version that hopefully will be available through open access by the end of 2022. Both books are accompanied by a website in github with colour versions of all graphs from the books, with added interactive features, and suggestions and R-scripts to replicate the measurements using the litRiddle R package.

The project was funded by the Computational Humanities Programme of the Royal Netherlands Academy of Arts and Sciences. It is currently being replicated in the United Kingdom. A follow-up project, The Riddle of the Literary Canon led by Karina van Dalen-Oskam, will analyse the style of older Dutch novels.

Bibliography

Eder, Maciej, Saskia Lensink, Joris van Zundert, Karina van Dalen-Oskam. The litRiddle package for R. CRAN, https://CRAN.R-project.org/package=litRiddle.

Koolen, C.W. 2018. *Reading Beyond the Female: The Relationship between Perception of Author Gender and Literary Quality.* PhD thesis, Amsterdam: University of Amsterdam. https://hdl.handle.net/11245.1/cb936704-8215-4f47-9013-0d43d37f1ce7

Koolen, Corina, Karina van Dalen-Oskam, Andreas van Cranenburgh, and Erica Nagelhout. 2020. 'Literary Quality in the Eye of the Dutch Reader: The National Reader Survey.' *Poetics* 79 (April): 101439. https://doi.org/10.1016/j.poetic.2020.101439.

Van Cranenburgh, Andreas. 2016. *Rich Statistical Parsing and Literary Language.* PhD thesis, Amsterdam: University of Amsterdam. http://andreasvc.github.io/phdthesis_v1.1.pdf.

Van Dalen-Oskam, Karina. 2021. *Het raadsel literatuur: Is literaire kwaliteit meetbaar?* [The riddle of literature. Can we measure literary quality?] Amsterdam: Amsterdam University Press. https://karinavdo.github.io/RaadselLiteratuur/

Van Dalen-Oskam, Karina. 2022. *The Riddle of Literary Quality. Measuring Perceptions of Literariness.* Amsterdam: Amsterdam University Press [in preparation].

Dehmel digital – Algorithmic-driven indexing of historical letters

Flüh, Marie
marie.flueh@uni-hamburg.de
Universität Hamburg, Germany

Bläß, Sandra
sandra.blaess@uni-hamburg.de
Universität Hamburg, Germany

On our poster, we present the *Dehmel digital* project (Hamburg University and Hamburg State and University Library; cf. Nantke, 2022) and its automatisation-oriented workflow for indexing historical letters. The project aims to digitise and index the letters of the correspondence network of Ida and Richard Dehmel, consisting of approximately 35,000 original handwritten letters. To be able to manage this huge amount of material, manual, semi-automatic and automatic/algorithm-driven work steps are interlinked in the project's workflow. Therefore, different machine learning techniques that belong to the methodological repertoire of the Digital Humanities are combined with each other. The goal is a digital network scholarly edition of the letters that makes the personal, cultural, and social dynamics contained in the correspondences tangible and explorable (cf. Nantke et al., 2022). There, the users can choose from various

medial perspectives on the documents, from facsimiles and transcriptions to registers and visualizations, to pursue their diverse (research) interests.

The workflow

The process of converting the handwritten originals into machine-readable representations is divided into a series of successive work stages that produce different data. Each result of the respective work steps is another layer of abstraction from the original material (see fig. 1).

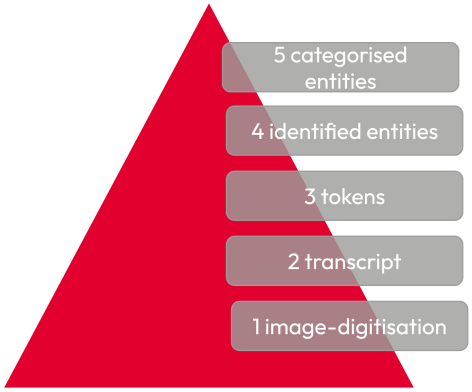


Figure 1:
Different results of the worklow

In order to establish a procedure that enables the indexing of such large manuscript corpora, but also keeps a philological standard of scholarly editions in mind, we combine several procedures already established within the Digital Humanities and modify them for the use in a digital edition (see fig. 2).

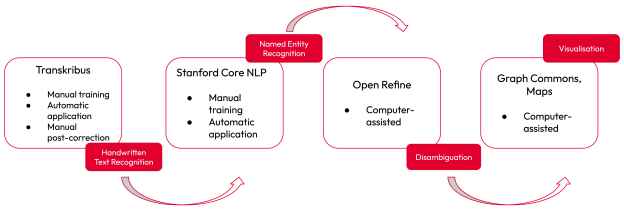


Figure 2:
The individual work steps in in the Dehmel digital project

After the original letters have been translated into high-resolution image digitisations and enriched with basic metadata, we transcribe about fifty handwritten document pages of one writer manually using Transkribus (cf. READ-COOP, 2021) until the data basis for training a Handwritten Text Recognition (HTR) model has been created. Following the successful training, further transcription is carried out in iterative alternation between automated transcription and manual post-correction, whereby the corrected letters are again fed into the HTR workflow as training data for a new model.

An XML processing chain converts these transcripts into a TEI-based format and applies a classifier of the Stanford Named Entity Recognizer (cf. Manning et al., 2014) to tag persons, places, institutions and artworks. The NER-classifier is adapted to the project's material base since it is trained manually to recognise these four different types of entities in letters from artists from the period around 1900. Recognised entities are introduced inline while retaining the references to the layout and structure of the document. For this purpose, a simplified implementation of the Separated Markup API for XML (cf. Verwer, 2020) is used (cf. Maus, 2021) which integrates classified tokens of NER and annotations of textual characteristics. The entities recognised in this way are used as the basis for the semi-automated generation of registers. These enlist all of the included people, places, institutions, and artworks and direct them back to the relevant documents. To convert the entities into stable register entries, a machine-supported data reconciliation is carried out with norm data systems and local knowledge bases. For reconciliation, we use OpenRefine (cf. OpenRefine, 2022) and the Gemeinsame Normdatei (GND) (cf. Deutsche Nationalbibliothek, 2022) as a norm data system.

The output

The computer-assisted disambiguation of entity types with OpenRefine (cf. OpenRefine, 2022) forms the basis for registers as well as for manually created macro comments on the central actors, institutions, and artworks of the corpus (see fig. 3).

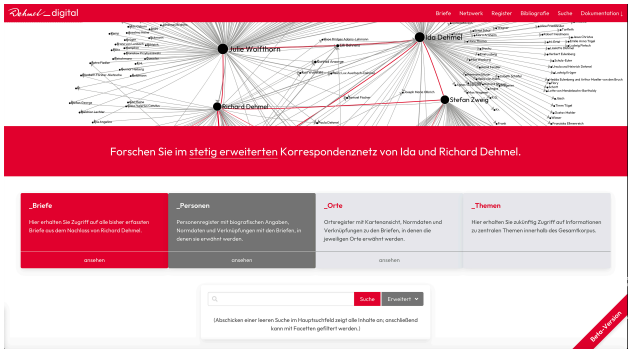


Figure 3:
The homepage of the Dehmel digital *project*

The algorithm-driven approach explained above ultimately enables both a basic overview of the entire corpus and partial correspondences. These are visualized as a dynamic, interactive network, which includes the letter's senders, recipients, as well as everyone mentioned within the letters, and directs back to the respective documents (see fig. 4). By linking central entities, visualisations, and digitised letters, it becomes possible to gain precise insights into the extensive corpus and use it as a scholarly edited historical source.

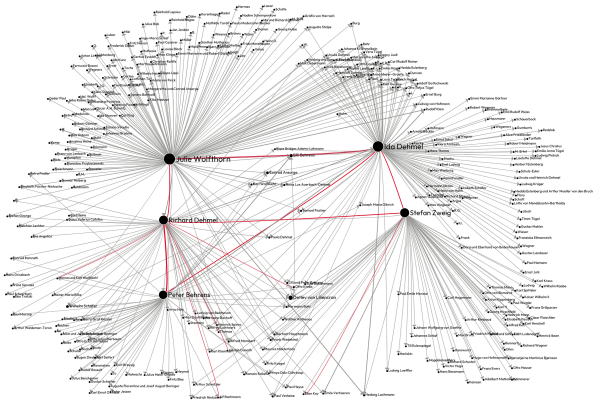


Figure 4:
Dynamic network visualisation of the correspondences as a network created with Graph Commons (cf. Arıkan et al., 2016)

Bibliography

Arıkan, B., Üstün, Z., Kızılay, A., Badur, A., Erikli, F., Zıngıl, Ö., Kılıçoğlu, D., Aldatmaz, A. and Dölec, G. (2016). Graph Commons. https://graphcommons.com (accessed 22 November 2021).

Deutsche Nationalbibliothek (2022): GND. Gemeinsame Normdatei. https://gnd.network/Webs/gnd/DE/Home/home_node.html (accessed 12 April 2022).

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J. and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, Maryland: ACL, pp.55–60.

Maus, D. (2021). XML NLP Pipeline, July 23, 2021. DOI: 10.25592/uhhfdm.9443.

Nantke, Julia (ed.) with the collaboration of Sandra Bläß and Marie Flüh (2022). Dehmel digital. https://dehmel-digital.de/ (accessed 19 April 2022).

Nantke, Julia, Sandra Bläß, and Marie Flüh (2022). Literatur als Praxis: Neue Perspektiven auf Brief-Korrespondenzen durch digitale Verfahren. In: Fischer, F. and Horstmann, J. (eds.): *Sonderband Text Praxis. Digitales Journal für Philologie. Digitale Verfahren in der Literaturwissenschaft.* DOI: https://doi.org/10.17879/64059432335.

OpenRefine (2022). OpenRefine. https://openrefine.org/ (accessed 22nd November 2021).

READ-COOP (2021). Transkribus. KI-gestützte Handschriftenerkennung. https://readcoop.eu/de/transkribus/ (accessed 22 November 2021).

Verwer, N. (2020). Plain text processing in structured documents. In: Proceedings of Declarative Amsterdam 2020. DOI: 10.1075/da.2020.verwer.plain-text-processing.

Quantitative Perspectives on European Baroque Drama: Towards a Network Theory-oriented Analysis

Giovannini, Luca
giovannini@uni-potsdam.de
Universität Potsdam, Germany

Introduction

The poster presents an ongoing doctoral research project which aims at exploiting the analytical power of network visualisation to investigate European Baroque drama from a quantitative perspective. By building and exploring a balanced corpus of 150 plays in five different languages, it attempts to provide an empirical verification of the traditional view on the evolution of seventeenth-century European dramatic literature.

Related literature

In the last two decades, network analysis of literary texts has established itself as a core methodology within the field of computational criticism and is now routinely employed to gain insights into social formations and character interactions within fictional worlds. Its strength lies in the extreme formalisation of texts, which are converted graphs made by characters (nodes) and their relations (edges); thus, it is possible to investigate large corpora and unearth formal patterns which close reading may overlook (cf. Trilcke, 2013; Trilcke and Fischer, 2018).

While many authors in the field have focused on methodological and technical issues, trying to leverage Natural Language Processing techniques to automate phases of the network extraction process, others have exploited the epistemological potential of literary networks analysis to try to answer questions concerning characters' features, textual topologies, and literary genres: a comprehensive overview on both approaches, with a wide bibliography, is provided by Labatut and Bost (2020). Furthermore, growing 'programmable corpora', with dedicated tools for network visualisation, are now available for research and teaching purposes (e.g. the Drama Corpora project by Fischer et al., 2019).

Project overview

This investigation of European drama is based on a research corpus of 150 plays, which is currently being assembled and will be later merged into the Drama Corpora repository (dracor.org). The corpus includes European plays in English, French, Spanish, German, and Italian, and covers the timespan from 1561 to 1710; despite its relatively small extension, its texts have been selected (or sampled from larger collections) with the explicit purpose of avoiding canonical bias and producing a composite and somehow 'representative' picture of the period investigated.

In a first phase, all plays which are not already available in DraCor are being transcribed or annotated to meet the platform's requirements, starting from structured or plain (.txt from OCRs) open-access textual sources and ending up with fully formatted XML-TEI files. Once established the corpus, character networks will be extracted from the texts by means of the DraCor scripts, which operate by linking characters by scene proximity, and visualised through Gephi (gephi.org) or similar software.

The textual structures embodied by the graphs will then be compared and interpreted according to the essential metrics of network analysis, such as centrality and modularity. The main aim will be measuring patterns of similarity and divergence between contemporary plays from different linguistic milieus and following the progressive evolution of drama throughout the designated temporal frame. Results are expected to contribute to the larger critical discussion on the features of Baroque, with a particular focus on the verification of some well-established literary theories the next section describes.

Research question and goals

The project is meant to address the lack of comparative studies on European early modern drama, and to complement the few existing ones (e. g. Küpper, 2018) with a 'quantitative formalist' perspective. Such shortage of extended transnational analyses of the genre has often been explained with the assumption that each 'local' form of Baroque drama represents a highly idiosyncratic system, sharply separated from the others by linguistic and cultural boundaries. This theory has been notably supported by Franco Moretti, who has described the evolution of European theatre throughout the seventeenth century as a process of Darwinist 'speciation'. In his view, indeed, the common heritage of classical and medieval drama progressively broke down into several national variations, such as the German Trauerspiel or the French théâtre classique, each one with his own set of distinctive stylistic and formal features (1994: 97-99).

Computational literary network analysis appears particularly suited to assess the validity of Moretti's reconstruction, since it is able to investigate the formal structures of all dramatic traditions involved with equal effectiveness. Accordingly, this study could help to measure whether (and how) Baroque-era theatre has split along national and cultural lines – and, conversely, to which extent phenomena of transfer of formal elements, such as plots or characters' roles, have nevertheless taken place. From this topological perspective, network-based evidence might thus contribute to a clearer understanding of the development of European dramatic literature during one of its defining periods.

Bibliography

Fischer, F., Börner, I., Göbel, M., Hechtl, A., Kittel, C., Milling, C. and Trilcke, P. (2019). Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. Proceedings of DH2019: 'Complexities'. Utrecht: Utrecht University doi:10.5281/ZENODO.4284002.

Küpper, J. (2018). The Cultural Net: Early Modern Drama as a Paradigm. Berlin and Boston: De Gruyter doi:10.1515/9783110536638.

Labatut, V. and Bost, X. (2020). Extraction and Analysis of Fictional Character Networks: A Survey. ACM Computing Surveys, 52 (5): 1–40 doi:10.1145/3344548.

Moretti, F. (1994). Modern European Literature: A Geographical Sketch. New Left Review (206): 86–109.

Trilcke, P. (2013). Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft. In Ajouri, P., Mellmann, K. and Rauen, C. (eds), Empirie in der Literaturwissenschaft. Münster: Brill | mentis, pp. 201–47 doi:10.30965/9783957439710_012.

Trilcke, P. and Fischer, F. (2018). Literaturwissenschaft als Hackathon. Zur Praxeologie der Digital Literary Studies und ihren epistemischen Dingen. In Huber, M. and Krämer, S. (eds), Wie Digitalität die Geisteswissenschaften verändert: Neue Forschungsgegenstände und Methoden. Wolfenbüttel: Herzog August Bibliothek doi:10.17175/sb003_003.

'Double blind' graph data analysis: a pedagogical experiment to discuss the intersubjectivity of network interpretation

Grandjean, Martin
martin.grandjean@unil.ch
History Department, University of Lausanne, Switzerland

Jacomy, Mathieu
mathieu.jacomy@gmail.com
TANTLab, Aalborg University in Copenhagen, Denmark

Preamble

Due to the hybrid nature of the conference and the uncertainty as to what the 'electronic posters' will ultimately be, we are intentionally proposing a format that takes advantage of the digital nature of the event: thus, the reflection around a means of presentation including interactive online elements and filmed content is an integral part of our approach. We are aware that this experiment goes beyond the framework of a standard conference paper, but we believe that the current pandemic must be an opportunity to rethink our remote collaboration and presentation formats.

Introduction

In this paper, we propose an experimental process of 'double blind' comparison which consists of giving the same dataset to two network analysts and to confront their methodological choices, results and interpretations. Recently tested in a format that combines recorded footage and live analysis (Jacomy and Grandjean, 2021), this 'double blind' approach makes it possible to make the intersubjectivity of the process visible and provide empirical material to discuss how the eyes of specialists from different disciplines influence the interpretation of the data. We believe that producing and making explicit such interpretive pathways can help researchers, designers and students understand the diversity of possible approaches.

The subjectivity of network interpretation in the humanities

If network analysis is a method now very widely diffused in the field of digital humanities (Ahnert et al., 2020), often because the metaphor and lexical field of 'network' are very effective in describing the objects of study of the humanities, there is no definitive interpretation procedure. Indeed, describing a network entails the recontextualization of graph-theoretical elements (e.g., the "betweenness centrality" metric) into the language of the discipline (e.g., "knowledge brokers" or simply "bridges") through a methodological "translation" (Grandjean and Jacomy, 2019). This hermeneutic process is subjective by nature, especially in the humanities where visual network analysis is often privileged, which may seem paradoxical considering that the method is supposed to objectify historical or literary sources (this is more generally true for all digital methods). Yet as Popper argued, descriptive statements necessarily draw their validity from

"intersubjective agreement" (Freeman, 1973). How much agreement does network analysis offer?

Experimental process

To ensure comparability, the phase of individual experimentation is framed by strict constraints in terms of time and tools. A network dataset is prepared upstream and duplicated: the two participants get a time budget to explore the data and produce an annotated network visualization of their choice, possibly including other tables and visualizations, as well as a list of hypotheses on the structure of the graph (fig. 1). This phase is filmed and screencasted. Then the two people discover each other's work and compare their interpretations.

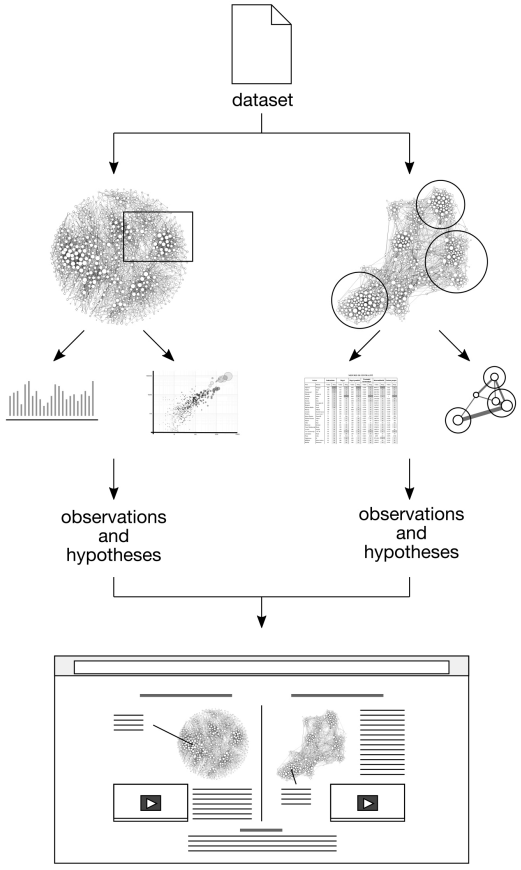


Figure 1.
Diagram of the process, the result is the window located at the bottom of the image, representing a web page comparing the interpretive paths by means of visualizations, texts and videos.

Results

The result of these parallel interpretations will consist of an interactive poster presenting the iterations of the two processes of analysis and interpretation. These two paths will be presented and compared by means of an online document integrating the data set, interactive and/or composite visualizations, descriptive text and video sequences making it possible to follow the reasoning. In the end, the data set itself is only a pretext for an educational experiment, the aim of which is as much to confront us with the subjectivity of the interpretation as to show the interest of such an exercise of replication.

Bibliography

Ahnert, R., Ahnert, S. E., Coleman, C. N. and Weingart, S. B. (2020). *The Network Turn: Changing Perspectives in the Humanities*. Cambridge: Cambridge University Press.

Jacomy, M. and Grandjean, M. (2021). Double Dating Data: Intellectual Cooperation in the League of Nations. Aarhus.

Freeman, E. (1973). Objectivity as "Intersubjective Agreement". *The Monist*, 57, 2: 168-175.

Grandjean, M. and Jacomy M. (2019). Translating Networks: Assessing Correspondence Between Network Visualisation and Analytics. *Digital Humanities 2019*. Utrecht.

Finding Ortese's Voice for Ferrante Fans: A Stylometric Study of *Neapolitan Chronicles*

Haggin, Patience
patiencehaggin@gmail.com
Independent Scholar, United States of America

Statement of purpose

When New Vessel Press published Anna Maria Ortese's *Neapolitan Chronicles* in 2018, it unabashedly marketed the book to fans of international best-selling author Elena Ferrante. Both authors were women who wrote fiction set in post-war Naples. The cover of *Neapolitan Chronicles* even

quoted Ferrante: "As for Naples, today I feel drawn above all by Anna Maria Ortese ..."

This paper stylometrically examines a translation of Anna Maria Ortese's 1953 *Il mare non bagna Napoli*. This paper concludes the translation was influenced by Ferrante's popularity.

Applicability

I intend for this paper to contribute to the growing body of research on translatorial style and to demonstrate that stylometric techniques originally developed to prove authorship can be used to analyse translatorial style (Rybicki, 2012; Rybicki, 2009; Lynch and Vogel, 2018; Forsyth and Lam, 2014).

Framework

My computational analysis relies on a code library created for stylometry in the programming language R by Drs. Maciej Eder, Mike Kestemont and Jan Rybicki (Eder et al., 107-121). This library, which was originally developed to resolve cases of disputed authorship, analyses authorial style by detecting patterns in each author's most frequent words. I present the results using methods that previous stylometric scholars have established as valid: scoring the texts based on stylometric similarity, comparing the most frequent words used by each author, and training a classifier to distinguish the authors' style (Rybicki et al., 123-144).

Methodology

Neapolitan Chronicles is a collaboration by Jenny McPhee and Ferrante's acclaimed translator Ann Goldstein. This paper investigates the nature of their collaboration and their claim to have "merged into yet another translator" with a unique, unified style (Goldstein and McPhee, 11). Stylometric study shows that they have indeed translated the book in a unified style, but that it is very similar to the style that Goldstein adopted for Ferrante translations.

This research relies on five corpora:

(1) *Neapolitan Chronicles*, translated by Ann Goldstein and Jenny McPhee

(2) Goldstein's translations of Ferrante novels, including the four novels that make
up her best-selling series

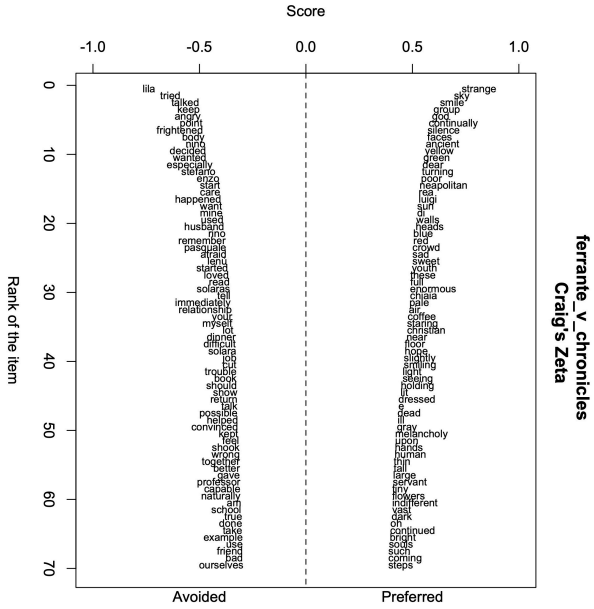
(3) English translations of Ortese's fiction by other translators

(4) Goldstein's solo fiction translations from Italian

(5) McPhee's solo fiction translations from Italian

Stylometric opposition

Below are the results from using stylometric opposition to compare *Neapolitan Chronicles* with Goldstein's translations of Ferrante's Neapolitan tetralogy. Stylometric opposition identifies the words that are "most unique" to each corpus.



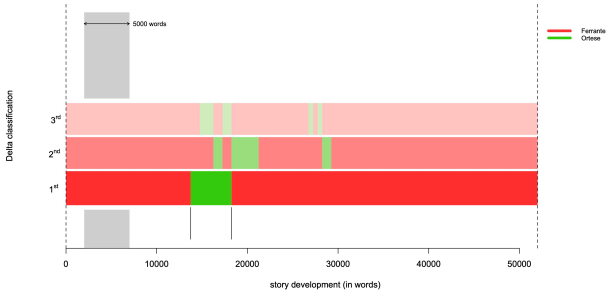
On the left are the 70 words most unique to Ferrante. On the right are the 70 words most unique to Ortese. Their distance to the right or left is based on the Zeta score of how distinctive they are. Many of Ferrante's most unique words are about school, family, or fear. Many of Ortese's most unique words come from her visual descriptions of a blighted city.

Notably, these lists are packed with "content words" (which address a book's topic) rather than "function words" (such as conjunctions, articles, prepositions and pronouns). Function words are a much better indicator of an author's signature style. Typically, when comparing two different authors, the results of stylometric opposition are dominated by function words. The results indicate these corpora use function words in very similar patterns—a classic sign of a deep similarity of writing styles.

Rolling classifier: Ferrante or Ortese?

Below are the results of a "rolling classifier" (Eder, 457-469) trained to recognize the most frequent words in Goldstein's Ferrante translations and other English

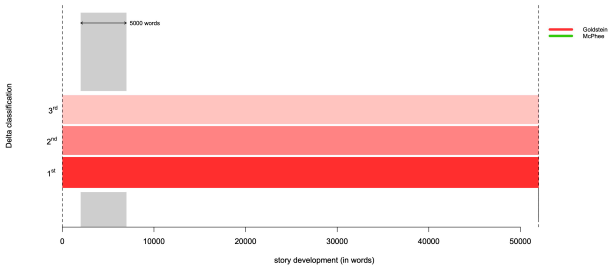
translations of Ortese. The classifier analyses *Neapolitan Chronicles* passage by passage and determines whether each segment is more similar to Goldstein's Ferrante or other translators' Ortese.



The horizontal band labeled "1st" gives the classifier's first judgment for each 5,000-word passage. The rolling classifier classified all of *Neapolitan Chronicles* as more stylometrically similar to Ferrante works, rather than Ortese works—with the exception of one story: "The Gold of Forcella," which focuses on the abject poverty of a Naples neighbourhood and has the most unique style.

Rolling classifier: Goldstein or McPhee?

Stylometry's rolling classifier has been proven capable of distinguishing individual translators' voices in cases of collaborative translation (Rybicki and Heydel, 708-717). Below are the results of a rolling classifier trained on Goldstein and McPhee's solo translations.



The classifier classified all parts of the book as more like Goldstein's work. The results support Goldstein and McPhee's claim to have "merged into yet another translator" with a coherent style throughout the book. Yet they don't support the claim that this translator had "a style all her own." Rather, this blended translator adopted a style very similar to the style that Goldstein used when translating Ferrante.

Stylistic features

This paper will discuss the stylistic features that *Neapolitan Chronicles* shares with Goldstein's Ferrante translations, such as: sentence length, gerunds, calqued Italian idioms and punctuation. This paper will discuss the stylistic traits that set "The Gold of Forcella" apart from the rest of *Neapolitan Chronicles*.

Conclusions

1. Goldstein and McPhee's collaboration achieved a unified style, yet their blended translator doesn't have "a style all her own." The pair adopted a style very similar to the style Goldstein used for Ferrante translations.

2. *Neapolitan Chronicles*' style is more similar to Goldstein's Ferrante than it is to any other English translations of Ortese. Goldstein and McPhee essentially "Ferrantized" Ortese's work to appeal to Ferrante fans. They subdued the unique style, which critics have described as "feverish," (Basile, 21) of Ortese's stories.

This paper is also a case study in the way that forgotten authors can see their reputations renewed and bolstered when contemporary authors that draw inspiration from them. This can be understood within David Damrosch's framework for understanding the hypercanon, countercanon and shadow canon, and the ways in which the three influence one another's status (Damrosch, 45-48). In this case, the renewed attention has influenced how a forgotten author is being translated and studied. Readers, publishers, and translators now read Ortese through a Ferrante lens.

Bibliography

Basile, E. (2014). *Anna Maria Ortese*. Ali&no editrice.

Damrosch, D. (2016). World Literature in a Postcanonical, Hypercanonical Age. *Comparative Literature in an Age of Globalization*. The Johns Hopkins University Press, pp. 43–53.

Eder, M. (2016). Rolling stylometry. *Digital Scholarship in the Humanities*, 31(3): 457–69.

Eder, M., Kestemont, M. and Rybicki, J. (2016). Stylometry with R: A package for computational text analysis. *R Journal*, 16(1): 107–21.

Forsyth, R. S. and Lam, P. W. Y. (2014). Found in translation: To what extent is authorial discriminability

preserved by translators?. *Literary and Linguistic Computing*, 29(2): 199–217.

Goldstein, A. and McPhee, J. (2018). Translators' Introduction. *Neapolitan Chronicles*. New Vessel Press, pp. 3–11.

Lynch, G. and Vogel, C. (2018). The translator's visibility: Detecting translatorial fingerprints in contemporaneous parallel translations. *Computer Speech & Language*, 52: 79–104.

Rybicki, J. (2009). Translation and Delta revisited: when we read translations, is it the author or the translator that we really read?. *Digital Humanities 2009: Conference Abstracts*. University of Maryland, College Park, pp. 245–47.

Rybicki, J. (2012). The great mystery of the (almost) invisible translator: stylometry in translation. *Quantitative Methods in Corpus-Based Translation Studies*. pp. 231–48.

Rybicki, J., Eder, M. and Hoover, D. (2016). Computational stylistics and text analysis. *Doing Digital Humanities*. Routledge, pp. 123–44.

Rybicki, J. and Heydel, M. (2013). The stylistics and stylometry of collaborative translation: Woolf's 'Night and Day' in Polish. *Literary and Linguistic Computing*, 28(4): 708–17.

Analysis of Exhibition Composition Using Co-occurrence Network Analysis

Hara, Shoko
shokohara@g.ecc.u-tokyo.ac.jp
The University of Tokyo, Japan

Ohmukai, Ikki
i2k@l.u-tokyo.ac.jp
The University of Tokyo, Japan

Nagasaki, Kiyonori
nagasaki@dhii.jp
International Institute for Digital Humanities, Japan

Takagi, Soichiro
stakagi@iii.u-tokyo.ac.jp
The University of Tokyo, Japan

We will focus on the difference in the contexts in which the artworks are placed in multiple exhibitions dealing with similar themes. This study targets the composition of special exhibitions in Japan and reveals that it is possible to clarify the difference in intentions of each work (Persohn, 2021) through visualization and comparison using a text-mining method called co-occurrence network analysis. An exhibition consists of some chapters, and elements that cross multiple chapters are important in understanding the intent of the whole exhibition. The subject of the analysis is the exhibitions on the theme of "Kyosai Kawanabe" held at three museums in central Tokyo in 2015, 2017, and 2019. Kyosai Kawanabe (1831-1889) was an ukiyo-e artist and Japanese-style painter active from the end of the Edo period to the beginning of the Meiji period.

The procedure for the analysis is as follows. First, the elements depicted in the work are extracted from the title of the works. They are all in Japanese and describe the motifs depicted. In them, the motifs are usually expressed as nouns or proper nouns, such as "a raven on a dead tree. Therefore, we used KH Coder (Higuchi, 2017), a kind of text mining tool, to extract specific parts of speech from the titles in each exhibition. In the above case, "raven" and "dead tree" are extracted, while "a" and "on" are eliminated. Next, we create a co-occurrence network diagram of chapter numbers and extracted words that appear two or more times for each exhibition. For each co-occurrence network, the number of occurrences of the extracted word is represented by the size of the circle, and the stronger the co-occurrence relationship, the darker the line. The numbers in the squares indicate the chapter numbers. In addition, the Jaccard coefficient, an indicator of the strength of the co-occurrence relationship between words, is also included for your information. The higher the coefficient, the more the motif is included in the chapter.

One of the features of this study is that we can see the range of interpretations that co-occurring words have, i.e., the many combinations of elements that are likely to be depicted in works placed in the same chapter. Although the outline of the exhibition can be obtained from the chapter titles, the co-occurrence network with the chapter number is the most effective visualization method to grasp the outline of the works placed in each chapter. When the co-occurrence of chapter and extracted words is known, it is easy to visually compare the exhibitions. It is also possible to compare the diversity of interpretations of the same word in different exhibitions.

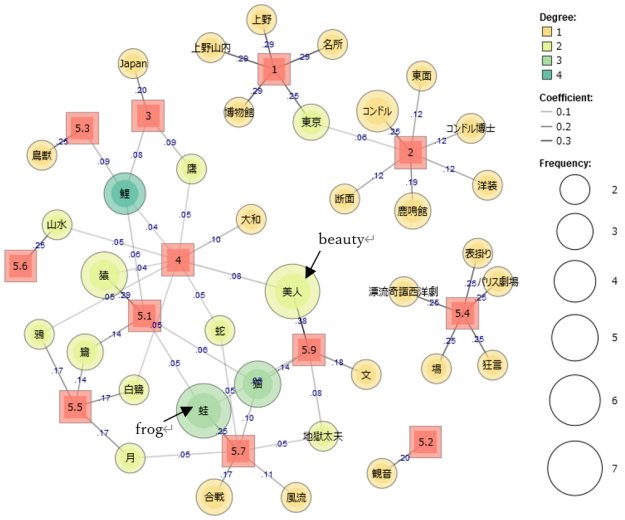


Figure1:
Exhibition in 2015.

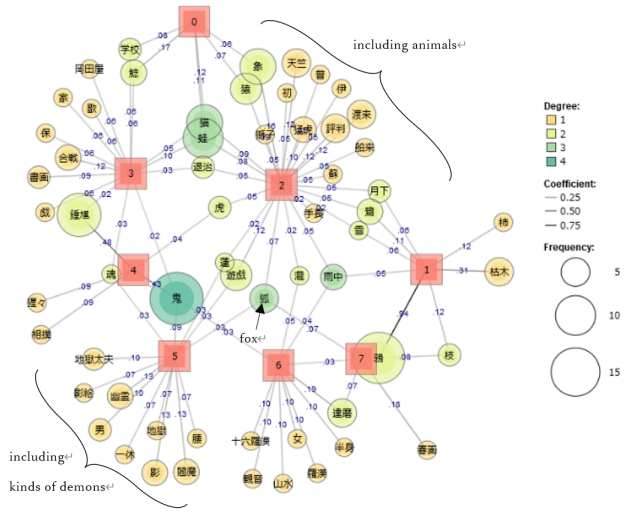


Figure2:
Exhibition in 2017.

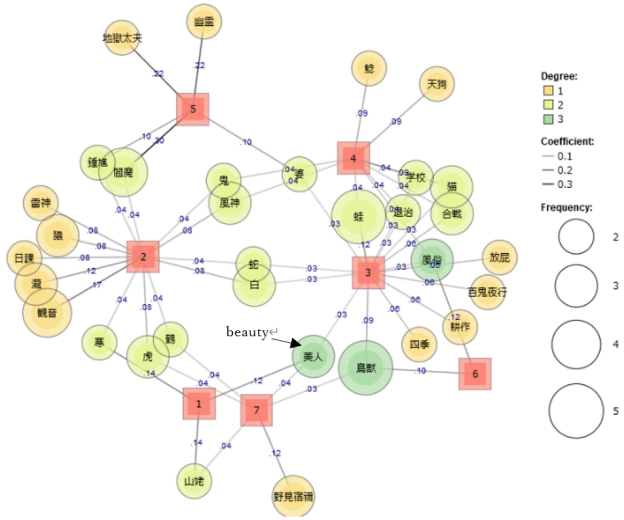


Figure3:
Exhibition in 2019.

As a result of our analysis, we have learned the following about each exhibition. For example, in the 2015 exhibition (Figure1), the most frequent occurrences "frog" and "beauty" are in co-occurrence relationship with chapter "4" or another number bigger than "5", i.e., works featuring frogs and beauties, which are prominent in number in the overall lineup, are placed in the latter half of the exhibition. In the 2017 exhibition (Figure2), the fox is described as an animal in the beginning, as a demon in the end, and as something that brings laughter in another chapter. The three exhibitions all share the fact that many of the works are related to animals and monsters, but the difference in the arrangement of the works lies in the intentions of the planners. The 2019 exhibition (Figure3) is the only one that begins and ends with “beauty”. It could be suggested that displaying artworks on a co-occurrence network makes it easier to grasp the content of an exhibition that is trying to capture artworks in terms of their techniques and styles.

This study has both academic and social significance in that it demonstrates the effectiveness of using the subject of exhibitions, which is difficult to analyze from a quantitative perspective, as a subject for mechanical analysis. Using this method and having the perspective of recognizing artworks from the perspective of chapter headings will make the experience of viewing artworks even more rewarding for individuals (Marty, 2011).

Bibliography

Persohn, L. (2021). Curation as methodology. Qualitative Research, 21(1): 20-41.

Higuchi, K. (2017). A Two-Step Approach to Quantitative Content Analysis: KH Coder Tutorial Using

Anne of Green Gables (Part II). Ritsumeikan Social Science Review, 53(1): 137-147.

Marty, P. F. (2011). My lost museum: User expectations and motivations for creating personal digital collections on museum websites. Library & information science research, 33(3): 211-219.

Development of datasets of the Hachidaishū and tools for the understanding of the characteristics and historical evolution of classical Japanese poetic vocabulary.

Hodošček, Bor
hodoscek.bor.hmt@osaka-u.ac.jp
Osaka University, Japan

Yamamoto, Hilofumi
yamamoto.h.al@m.titech.ac.jp
Tokyo Institute of Technology, Japan

The purpose of this study is to update the Hachidaishu (ca. 905–1205; about 9500 poems) dataset (Yamamoto and Hodošček 2021a, 2021b), which is already available on Zenodo/Github. We provide details on changes to 1) the data description format, 2) the Word List by Semantic Principles' labels (Kato et al. 2018), and 3) updates and item additions to the analysis program.

The Hachidaishū is a collection of eight anthologies of classical Japanese poetry compiled by the order of Emperors during the 300 years from the Kokinshū (ca. 905), the first anthology written in Japanese kana characters, to the Shinkokinshū (1205). The main text is based on a collection of the Nijūichidaishū created by NIJL, and the text is now distributed by both NIJL and NII (National Institute of Japanese Literature 2016). Classical Japanese poetry is not only a literary work created through singing, which is of great literary value in itself, but also a work created based on the phonology of the spoken language of the time, which is an extremely valuable research material for linguistic studies of classical Japanese.

The Japanese imperial anthologies of classical Japanese poetry (waka: wa=Japanese, ka=song) were composed over a period of 300 years using the same form and with the same divisions, called 'butate', allowing us to compare waka poems with each other. These anthologies are convenient for analyzing the historical changes in Japanese language. We have published the datasets on Zenodo and Github, which allow Japanese linguists and literary researchers to study them on the data analysis basis (Yamamoto and Hodošček 2021c). However, because the format used in these datasets was original, they were not generic enough for comparisons with other works or for use with publicly available tools. In order to solve these problems, we updated the datasets. The updates and additions include the following items:

1. Replacing the current fixed-length data description format with TEI and JSON, which are general-purpose data description formats, so that they can be used by more analysis tools. The purpose of the JSON conversion is to provide a more self-describing and accessible version of the data, while that of the TEI conversion is to have a standard version of the dataset as a corpus going forward.
2. Replaced the Word List by Semantic Principles' labels (WLSP; Kato et al. 2018) from the old version to the new version to allow us to compare them with other datasets.
3. The part-of-speech (PoS) tags were previously based on the PoS numbers used in an older Japanese morphological analysis system, so we added a mapping to the Universal Dependencies tag set (17 types). This will allow us to compare lexical items, syntactic rules, and structures among different languages and works.
4. We have developed tools that can be freely used on researchers' hardware and also provide a hosted version on Google Collaborator to allow other researchers to use and analyze these datasets. For example, basic statistics, conditional string search, and visualization are available. To these tools, we have added programs that can draw graphs using bi-gram patterns, so that the structure of the vocabulary can be easily visualized.
5. Since the general lexical set does not include proper nouns (names of people, places, etc.), we included data on pillow names and place names included in the Hachidaishū.
6. An English bilingual dataset was created, allowing for easy generation and output of English glosses when presenting at international conferences or writing international papers.

These updates will contribute to the elucidation of the characteristics and historical evolution of Japanese poetic vocabulary in more detail.

Bibliography

Kato S., Asahara M., and Yamazaki, M. (2018) "Annotation of 'Word List by Semantic Principles'

Labels for the Balanced Corpus of Contemporary Written Japanese", in Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation, Hong Kong: Association for Computational Linguistics, pp. 247-253.

Yamamoto, H. and Hodošček B. (2021a) "Hachidaishū part of speech dataset", https://doi.org/10.5281/zenodo.4835806.

Yamamoto, H. and Hodošček B. (2021b) "Hachidaishū vocabulary dataset", https://doi.org/10.5281/zenodo.4744170.

Yamamoto, H. and Hodošček B. (2021c) "Open source datasets of the Hachidaishū for the research of classical Japanese poetic vocabulary", The 11th Conference of Japanese Association for Digital Humanities, Proceedings of JADH conference, Japanese Association for Digital Humanities, Vol. 2021, pp. 82-87, Sept.

Development of a *Devanāgarī* Optical Character Recognition (OCR) System

Kato, Takahiro
tkhrkt@l.u-tokyo.ac.jp
The University of Tokyo, Japan

Tomonari, Yūki
mitriibhaavii@gmail.com
The University of Tokyo, Japan

Taniguchi, Chikamitsu
tanigutiti@g.ecc.u-tokyo.ac.jp
The University of Tokyo, Japan

Osawa, Tomejiro
tomejiro.osawa@toppan.co.jp
Toppan Inc.

Fujimaki, Satoshi
satoshi.fujimaki@toppan.co.jp
Toppan Inc.

Okada, Takashi
takashi_3.okada@toppan.co.jp
Toppan Inc.

Hashimoto, Emi
emi_1.hashimoto@toppan.co.jp
Toppan Inc.

This poster outlines the objectives of the research project titled "Development of a *Devanāgarī* Optical Character Recognition (OCR) System."

Devanāgarī is an abugida script, which has been adopted as the writing system of several languages such as Hindī, Marāṭhī, Nepālī, and Sanskrit. Recently, digitizing Sanskrit texts written in *Devanāgarī* has been one of the most pressing and important tasks in the field of Sanskrit philology. Prosopographical Database for Indic Texts (PANDiT), Sanskrit Knowledge-System Project, and Göttingen Register of Electronic Text in Indian Languages (GRETIL) are some of the leading research projects.

However, owing to the costs associated with time and human labor, building a database based on manually input text data is challenging. We have seen this in some preceding projects led by scholars in Germany, India, and Japan. We expect that the existing *Devanāgarī* OCR systems, developed often based on the contemporary Indic languages such as Hindī ([1][2]), may not accurately recognize the more complicated Sanskrit consonant clusters.

In light of this situation, a team comprising Sanskrit language experts from the University of Tokyo and AI-OCR developers from Toppan Inc. have undertaken a cooperative research project. This project aims to develop a *Devanāgarī* OCR system and establish a Sanskrit text database automatically digitized by the OCR.

In this poster presentation, we will

- Review the writing system of *Devanāgārī* and describe how we correlate each combining letter with the Unicode encoding scheme. We took each letter as a composite of several elements. In this case, our experience of the Chinese character—a character consisting of multiple and irregularly ordered elements —served effectively. In this regard, we set a unit of letter called the "character shape."
- Introduce and evaluate preceding and on-going OCR software such as the "Sanskrit OCR" run by ind.senz and "Google Document OCR." According to our detailed analysis of these software programs, there are some specific cases where these OCRs frequently fail to recognize the letters. For example: some combined letters with the vowel sign "*i* (ि)," where the sequence of letter elements (right to left, e.g. *k*+*i*) goes against the stroke order (left to right, e. g. ि+क, *i*+*k*); some irregularly typeset dots, which indicate the nasal sound (*anusvāra*), and some lengthy consonant clusters such as (त्स्न्र्य, *rtsny-a*). Focusing on these inadequacies that were insufficiently handled by the preceding studies, we show our design of an AI-OCR model, highlighting the uniqueness of this project.

- Expound the process of designing the “training data” through which an AI-OCR is generated. We obtained the training data from books included in Ānanda Āśrama Sanskrit Series, most of which were printed in metal typesetting. The strategy on how we define the “character shape” in the typesetting shall be explained in detail. An AI-OCR was generated through machine learning using the datasets prepared through the above process. Following is a brief overview of the outcomes obtained from the generated AI-OCR model.

Outcomes of Single Character Recognition (as of February 15, 2022)

Out of the 2,434 sample letters:

a. 2,340 letters exactly recognized* (Accuracy rate 96.14 %)

b. 2,397 letters correctly listed** (Accuracy rate 98.48 %)

* Only when each letter is listed as the first choice.

** Including cases where the correct letter is listed as a candidate.

Based on the comparison, "Google Document OCR" and "Sanskrit OCR" showed an accuracy rate of 95.00 % and 89.92 %, respectively, on average for the common samples in this presentation. Once the factors affecting the accuracy are understood, we will outline the adjustments needed to improve the accuracy rate of the AI-OCR.

Bibliography

Bansal, V and Sinha, M. (2001). A Complete OCR for Printed Hindi Text in Devanagari Script. *Proceedings of Sixth International Conference on Document Analysis and Recognition*, pp. 800-804.

Suryaprakash Kompalli, Sankalp Nayak, Srirangaraj Setlur and Venu Govindaraju (2005). Challenges in OCR of Devanagari documents. *Eighth International Conference on Document Analysis and Recognition (ICDAR'05),* pp. 327-331.

Analysis and Exploration of Supernatural Fanfictions from the Platform Archive of Our Own

Kleindienst, Nina
Nina.Kleindienst@stud.uni-regensburg.de
Media Informatics Group, University of Regensburg, Germany

Schmidt, Thomas
thomas.schmidt@ur.de
Media Informatics Group, University of Regensburg, Germany

Wolff, Christian
christian.wolff@ur.de
Media Informatics Group, University of Regensburg, Germany

Introduction

Fanfictions are fan-created written works using already existing characters and plot elements of existing famous media to write new stories based on those characters (Dym et al. 2018). This genre of online literature has gained a lot of interest in literary studies and the humanities (cf. Hellekson and Busse, 2006; Thomas, 2011; Van Steenhuyse, 2011; Jamison, 2013) but also in digital humanities (DH) in recent years (Milli & Bamman, 2016; Fast et al. 2016; Yin et al. 2017; Frens et al. 2018; Rebora and Pianzola, 2018; Pianzola et al. 2020; Schmidt et al. 2021d).

We present first results of the analysis of around 15 years of fanfiction production for the fandom Supernatural and fanfictions from the platform Archive of Our Own (AO3) [1]. Supernatural is a popular mystery-fantasy American TV show running from 2005 to 2020 and the fan fiction community is regarded as one of the most productive which has led to research in DH (Kleindienst and Schmidt, 2020). AO3 is one of the most popular fan fiction websites and hosts, to our knowledge, the largest number of Supernatural fan fictions compared to other platforms.

Corpus

We developed a script to scrape all Supernatural fanfictions from AO3 which we used in January 2021 to gather all fan fictions available at that moment. [2] We scraped the entire HTML page and transformed the content into structured JSON including the text of the stories and metadata. The general corpus analysis was performed similar to research on social media (e.g. Moßburger et al., 2020) Overall, the corpus consists of 170,436 unique fan fictions and over 1 billion tokens. On average a fanfiction has a length of around 6,000 tokens but this length varies between 100 to 2 million tokens (table 1).

Metric	Value
Number of unique fanfictions	170,436
Size of corpus (measured as HTML-files)	13.82 GB
Examples for metadata per fanfiction	e.g., Fandom, Relationships, Kudos
Number of tokens (only for the text of the fanfictions)	1,059,259,740
Average number of tokens per fanfiction	6,215.7
Average number of sentences per fanfiction	406.1

Table 1.
General corpus statistics.

Results

In the following section, we present a small subset of first corpus and metadata results. We focus on diachronic analysis across the respective time spans of the seasons of the show from 2005-2020. Each season basically represents one year. We regard fanfictions as part of a season if the publication date is during the airing of this seasons or before the airing of the next season.

As figure 1 shows, the most significant increase in production begins in season 4 and peaks in season 10. Indeed, season 4 sees the appearance of the character "Castiel" which is of great importance for the community.

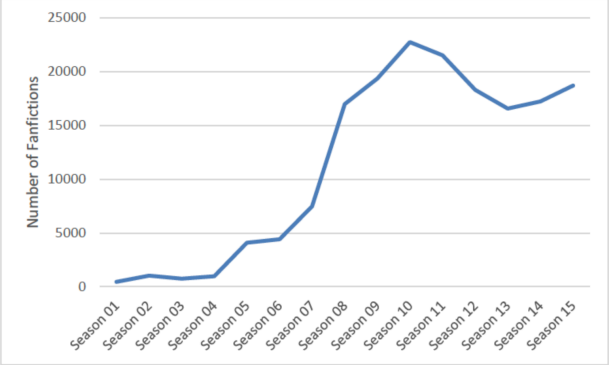


Figure 1.
Fanfiction production across seasons.

We also compared the production of the airing time of a season to the in-between time. As figure 2 shows, the production takes a significant decrease in between seasons. This is in line with research by De Kosnik et al. (2015) who have found that fanfiction production tends to happen immediately after release.

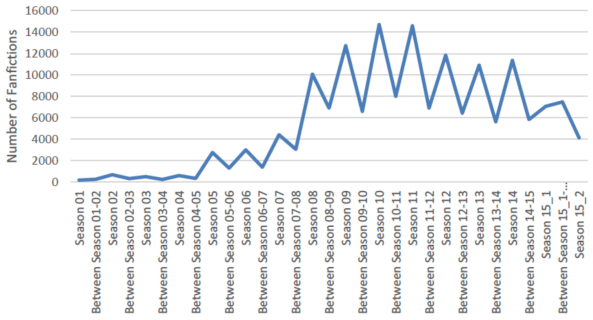


Figure 2.
Fanfiction production across and in-between seasons.

The average length of the individual fanfiction increases throughout the airing of the show peaking with almost 8,000 tokens in the more recent seasons (figure 3).

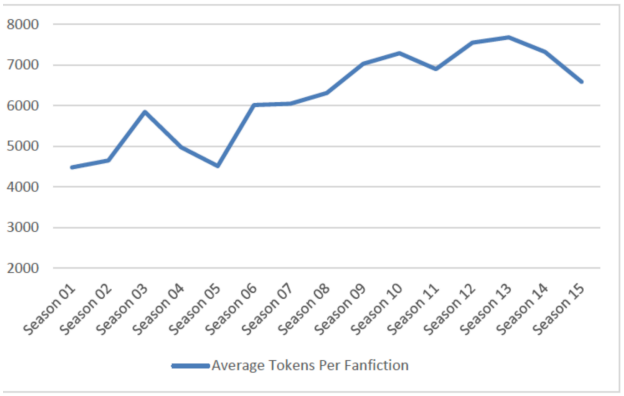


Figure 3.
Average number of tokens per fanfiction.

Considering the number of different authors, we found that up until season 6 this number stays rather small below a limited core of 1,000 authors but increases drastically, again, in season 4 and season 7 (figure 4).

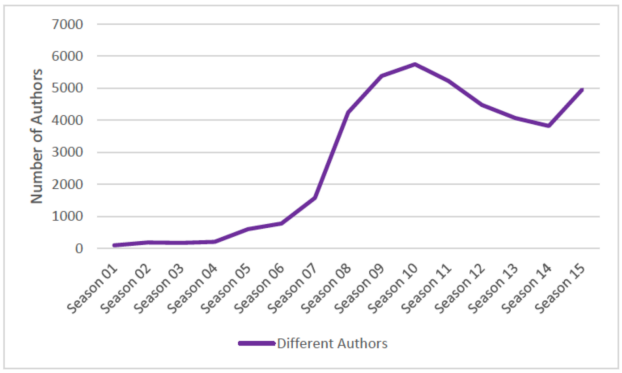


Figure 4.
Number of authors per season.

One important factor of fanfictions is the relationship type of the characters the fanfiction depicts. Indeed, as the overall distribution of this AO3 metadata shows, the majority (55.4%) of all fanfictions of this corpus deals with male-male homo-erotic and romantic content (slash). This is a well-known phenomenon of fanfictions (cf. Hellekson & Busse, 2006) and an important part of Supernatural fanfictions.

Category	Percentage of Fanfictions
M/M	55.4%
F/M	14.9%
F/F	1.6%
Gen	28.4%
Multi	2.6%
Other	1.9%
No Category	4.7%

Table 16: Category Analysis (Fanfiction Corpus)

Table 2.
Distribution of relationship type tags for the entire corpus.

Again, the dominance of this slash fanfictions (M/M) becomes striking over time beginning in season 4 as figure 5 shows.

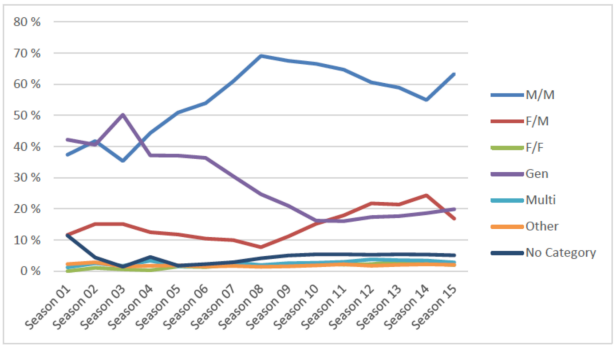


Figure 5.
Proportion of relationship type tags across seasons.

Fanfiction authors can explicitly add the specific character relationship as metadata which is also curated by AO3. Looking at the proportion of the three most popular relationships (figure 6), we identified that (1) all of them are of M/M-nature and (2) the rise in popularity of the show and the fanfictions goes hand in hand with the introduction of the character Castiel and the imaginations about the relationship with the main character Dean Winchester. This is especially striking since Castiel is a mere side character in season 4 that became part of the main cast due to his popularity that can also be seen in our data.

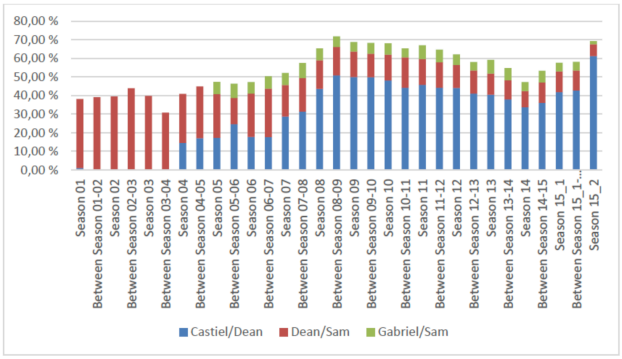


Figure 6.
Proportion of the three most popular relationships across seasons.

Please note, that we only presented a subset of the results this corpus has to offer. Furthermore, we also see great potential form more advanced methods that have gained popularity in recent years in DH like sentiment analysis (Schmidt and Burghardt, 2018; Schmidt et al., 2021a; Schmidt et al., 2021b) or even multimodal approaches including the video channel of the TV show in further studies (similar to Schmidt et al., 2019; Schmidt et al., 2020a; Schmidt et al., 2020b; Schmidt et al., 2021c; Schmidt and Wolff, 2021).

Bibliography

De Kosnik, A., El Ghaoui, L., Cuntz-Leng, V., Godbehere, A., Horbinski, A., Hutz, A., Pastel, R. and Pham, V. (2015). Watching, creating, and archiving: Observations on the quantity and temporality of fannish productivity in online fan fiction archives. *Convergence*, **21**(1). SAGE Publications Ltd: 145–64 doi: 10.1177/1354856514560313.

Dym, B., Aragon, C., Bullard, J., Davis, R. and Fiesler, C. (2018). Online Fandom: Boldly Going Where Few CSCW Researchers Have Gone Before. *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*. Jersey City NJ USA: ACM, pp. 121–24 doi: 10.1145/3272973.3274542. https://dl.acm.org/doi/10.1145/3272973.3274542 (accessed 6 October 2020).

Fast, E., Vachovsky, T. and Bernstein, M. S. (2016). Shirtless and Dangerous: Quantifying Linguistic Signals of Gender Bias in an Online Fiction Writing Community. *ICWSM*.

Hellekson, K. and Busse, K. (2006). *Fan Fiction and Fan Communities in the Age of the Internet: New Essays*. McFarland.

Jamison, A. (2013). *Fic: Why Fanfiction Is Taking Over the World*. Illustrated Auflage. Dallas, Texas: Smart Pop.

Kleindienst, N. and Schmidt, T. (2020). Investigating the Transformation of Original Work by the Online Fan Fiction Community: A Case Study for Supernatural. Basel, Switzerland https://digitalpractices2020.philhist.unibas.ch/en/abstracts/#ABSTRACTSCHMIDT (accessed 21 April 2022).

Milli, S. and Bamman, D. (2016). Beyond Canonical Texts: A Computational Analysis of Fanfiction. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 2048–53 doi: 10.18653/v1/D16-1218. http://aclweb.org/anthology/D16-1218 (accessed 6 October 2020).

Moßburger, L., Wende, F., Brinkmann, K. and Schmidt, T. (2020). Exploring Online Depression Forums via Text Mining: A Comparison of Reddit and a Curated Online Forum. *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*. Barcelona, Spain (Online): Association for Computational Linguistics, pp. 70–81 https://aclanthology.org/2020.smm4h-1.11 (accessed 21 April 2022).

Pianzola, F., Rebora, S. and Lauer, G. (2020). Wattpad as a resource for literary studies. Quantitative and qualitative examples of the importance of digital social reading and readers' comments in the margins. (Ed.) Orrego-Carmona, D. *PLOS ONE*, **15**(1): e0226708 doi: 10.1371/journal.pone.0226708.

Rebora, S. and Pianzola, F. (2018). A New Research Programme for Reading Research: Analysing Comments in the Margins on Wattpad. *DigitCult | Scientific Journal on Digital Cultures*(3.2): 19–36 doi: 10.4399/97888255181532.

Schmidt, T. and Burghardt, M. (2018). An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing. *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Santa Fe, New Mexico: Association for Computational Linguistics, pp. 139–49 https://www.aclweb.org/anthology/W18-4516 (accessed 6 April 2020).

Schmidt, T., Burghardt, M. and Wolff, C. (2019). Toward Multimodal Sentiment Analysis of Historic Plays: A Case Study with Text and Audio for Lessing's Emilia Galotti. In Navarretta, C., Agirrezabal, M. and Maegaard, B. (eds), *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference (DHN 2019)*. Copenhagen, Denmark, pp. 405–14 http://ceur-ws.org/Vol-2364/37_paper.pdf (accessed 21 April 2022).

Schmidt, T., Dangel, J. and Wolff, C. (2021a). SentText: A Tool for Lexicon-based Sentiment Analysis in Digital Humanities. vol. 74. Glückstadt: Werner Hülsbusch, pp. 156–72 https://epub.uni-regensburg.de/44943/ (accessed 21 April 2022).

Schmidt, T., Dennerlein, K. and Wolff, C. (2021b). Emotion Classification in German Plays with Transformer-based Language Models Pretrained on Historical and Contemporary Language. *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Punta Cana, Dominican Republic (online): Association for Computational Linguistics, pp. 67–79 doi: 10.18653/v1/2021.latechclfl-1.8. https://aclanthology.org/2021.latechclfl-1.8 (accessed 21 April 2022).

Schmidt, T., El-Keilany, A., Eger, J. and Kurek, S. (2021c). Exploring Computer Vision for Film Analysis: A Case Study for Five Canonical Movies. *2nd International Conference of the European Association for Digital Humanities (EADH 2021)*. Krasnoyarsk, Russia https://www.researchgate.net/publication/355446064_Exploring_Computer_Vision_for_Film_Analysis_A_Case_Study_for_Five_Canonical_Movies (accessed 21 April 2022).

Schmidt, T., Engl, I., Halbhuber, D. and Wolff, C. (2020a). Comparing Live Sentiment Annotation of Movies via Arduino and a Slider with Textual Annotation of Subtitles. In Reinsone, S., Skadiņa, I., Daugavietis, J. and Baklāne, A. (eds), *Post-Proceedings of the 5th Conference Digital Humanities in the Nordic Countries (DHN 2020)*, vol. 2865. Riga, Latvia: CEUR Workshop Proceedings, pp. 212–23 http://ceur-ws.org/Vol-2865/poster1.pdf (accessed 21 April 2022).

Schmidt, T., Grünler, J., Schönwerth, N. and Wolff, C. (2021d). Towards the Analysis of Fan Fictions in German Language: Exploration of a Corpus from the Platform Archive of Our Own. Krasnoyarsk, Russia https://www.researchgate.net/publication/355228481_Towards_the_Analysis_of_Fan_Fictions_in_German_Language_Exploration_of_a_Corpus_from_the_Platform_Archive_of_Our_Own (accessed 21 April 2022).

Schmidt, T., Mosiienko, A., Faber, R., Herzog, J. and Wolff, C. (2020b). Utilizing HTML-analysis and computer vision on a corpus of website screenshots to investigate design developments on the web. *Proceedings of the Association for Information Science and Technology*, **57**(1): e392 doi: 10.1002/pra2.392.

Schmidt, T. and Wolff, C. (2021). Exploring Multimodal Sentiment Analysis in Plays: A Case Study for a Theater Recording of Emilia Galotti. *Proceedings of the Conference on Computational Humanities Research 2021 (CHR 2021)*. Amsterdam, The Netherlands, pp. 392–404.

Thomas, B. (2011). What Is Fanfiction and Why Are People Saying Such Nice Things about It?. *Storyworlds: A Journal of Narrative Studies*, **3** doi: 10.5250/storyworlds.3.2011.0001.

Van Steenhuyse, V. (2011). The Writing and Reading of Fan Fiction and Transformation Theory. *CLCWeb: Comparative Literature and Culture*, **13**(4) doi: 10.7771/1481-4374.1691. https://docs.lib.purdue.edu/clcweb/vol13/iss4/4 (accessed 7 October 2020).

Yin, K., Aragon, C., Evans, S. and Davis, K. (2017). Where No One Has Gone Before: A Meta-Dataset of the World's Largest Fanfiction Repository. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Denver Colorado USA: ACM, pp. 6106–10 doi: 10.1145/3025453.3025720. https://dl.acm.org/doi/10.1145/3025453.3025720 (accessed 6 October 2020).

Notes

1. https://archiveofourown.org/
2. Due to legal constraints, the data is only available via request (thomas.schmidt@ur.de).

Répertoire des Écritures Numériques : archiving and qualifying electronic literature

Lescouet, Emmanuelle
emmanuelle.lescouet@umontreal.ca
Université de Montréal, Canada

Vitali-Rosati, Marcello
marcello.vitali.rosati@umontreal.ca
Université de Montréal, Canada

The *Répertoire des Écritures Numériques* is a platform created in 2015 by the Canada Research Chair in Digital Textualities, under the direction of Marcello Vitali-Rosati, and which later joined the inter-university partnership Littérature Québécoise Mobile. It has undergone many redesigns and transformations, moving from the Repertoire of Digital Writers (Vitali-Rosati, 2017), focused on figures of auctoriality, to become the Répertoire it is today, interested in the real and fluid practices of literature. This project is now lead by Emmanuelle Lescouet.

This transition has become necessary with the increase of literary artworks on various media and platforms, including game consoles and smartphones, connected watches or immersive worlds (Ryan, 2015). The place of the author is fading away to make room for fluid collectives, entities and representations of the auctorial and editorial function. The literary tradition leads us to multiple art forms, inscribed into different supports of reading. This actually brings us to consider new receptions and actual interaction with the text.

The importance of the inscription in the traditions and digital cultures of the different communities, so much related to the supports of reading. We have to emphasize the production and the concrete means of the existence of the artworks to identify possible receptions.

The trend towards intermediality, through the approach of videogame narratives (Aarseth, 1999; Hayles, 2007; Gervais and Archibald, 2006), network theories and practices (Manovich 2001), leads us to build a highly heterogeneous corpus place (Emerit, 2016). The selection of artworks raises questions, such as which ones could enter the 'digital literature'? What are we to do with hybrid propositions? We are compulse to think literature in its inscription and its interactions with other close artforms such as video games or interactive movies. The approach here is not to legitimize artworks that exist by themselves outside the hands of academia, but to allow the documentation of a practice.

Historical studies on digital literary corpora, based in part on hypertext and arborescent organizations, call for being surpassed in an attempt to approach a contemporary theory of digital reading.

As digital literary works adapt to different media and techniques in perpetual and fast evolution, there is a research challenge to document these works, before their disappearance or their shift in practices. To capture them, or at least to gather enough information on them and the practices that accompany them to be able to draw up statistical studies and analyses of reading and its contemporary forms.

To this end, the *Répertoire* aims to document and catalog the emerging forms of digital literature, which leads the team to question the possible evolutions of artforms and so of categorizations and denominations to best qualify them. The establishment of characteristic fields, describing the concrete and formal perceptions of the work, is necessary for a documentation of this corpus. This documentation is then enriched through a literary approach, calling upon the narrative form as well as the discursive genre, the themes or the organization of the text.

The *Répertoire* refers to a double approach: in one hand it analyses the technology, in the other it focuses on the reader body's implications by the reading moves set up. All this is happening thanks to Omeka S, combined with a thesaurus via OpenTheso. The link between the two enrich the database and its implementation in the records allows a more precise description, validated by the community. The construction of a collaborative and peer-reviewed ontology within academia provides stability and a shared definition of concepts. Establishing common names for emerging literary

forms allows us to document them and their evolution to be identified.

Understanding how these artworks are constructed – what kind of software or content management systems are used – makes it possible to identify the technological constraints with which they are created; their possible reading hardwares as well as the community history of the studied practices. When linked to the medium, the technology leads us to consider the possible interactions with the artwork: from naming and combining gesture taxonomies to the physical interface between the reader and the artwork (Galloway, 2011; Garmon 2018; Souchier et al. 2019).

The implication of the body and its semantic construction materialize this incarnation and project the reader in a particular reception: if the artwork is composed by already acquired gestures, well-known interfaces, or foreign environment, the immediacy of the experience will place the reader in a different immmersive state. The digital environment, as well as our hyperconnected daily practices, allows us to understand the place of the literary artworks in the daily life of each one.

This poster will present a visual representation of the collected data. Through this cartography we will explain the necessity of the various categories and their inter-action into the reception and the formal offered experiment.

Bibliography

Aarseth, E. (1997) *Cybertext: Perspectives on Ergodic Literature*. Baltimore: Johns Hopkins University Press.

Bouchardon, S. (2009) *Littérature numérique : le récit interactif*. Paris: Hermès Science.

Bouchardon, S. (2012) 'Du récit hypertextuel au récit interactif', *Revue de la BNF*, (42), pp. 13–20.

Emerit, L. (2016) 'La notion de lieu de corpus : un nouvel outil pour l'étude des terrains numériques en linguistique', *Corela*, 14(1). doi:https://doi.org/10.4000/corela.4594.

Galloway, A.R. (2012) *The interface effect*. Cambridge, UK ; Malden, MA: Polity.

Garmon, I. (2020) 'Le corps à l'épreuve des applications : des « petits gestes » éprouvants ?', *Les Chantiers de la Création*, La mise à l'épreuve du corps(12). doi:https://doi.org/10.4000/lcc.3102.

Gervais, B. and Archibald, S. (2006) 'Le récit en jeu : narrativité et interactivité', *Protée*, 34(2–3), pp. 27–29.

Hayles, K.N. (2007) *Electronic literature: what is it?* Available at: https://eliterature.org/pad/elp.html (Accessed: 2 April 2019).

Manovich, L. (2001) *The language of new media*. Cambridge, Mass.: MIT Press (Leonardo).

Ryan, M.-L. (2015) *Narrative as Virtual Reality 2*. Baltimore: Johns Hopkins UP.

Souchier, E., Candel, É. and Gomez-Mejia, G. (2019) 'Regards sur le numérique', in *Le numérique comme écriture*. Paris: Armand Colin (Théories et méthodes d'analyse), pp. 21–191.

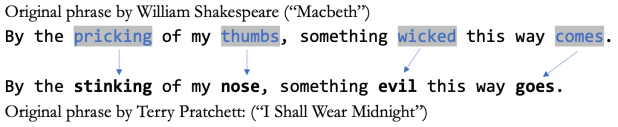
Vitali-Rosati, M. (2018) *La littérature numérique francophone : enjeux théoriques et pratiques pour l'identification d'un corpus*, *Culture numérique. Pour une philosophie du numérique*. Available at: http://blog.sens-public.org/marcellovitalirosati/la-litterature-numerique-francophone-enjeux-theoriques-et-pratiques-pour-lidentification-dun-corpus/ (Accessed: 22 November 2021).

The Vectorian API – A Research Framework for Semantic Textual Similarity (STS) Searches

Liebl, Bernhard
liebl@informatik.uni-leipzig.de
Leipzig University, Germany

Burghardt, Manuel
burghardt@informatik.uni-leipzig.de
Leipzig University, Germany

In the humanities, texts are often quoted, referenced or alluded to (see Bamman & Crane, 2008). In order to automatically detect complex cases of so-called intertextual references, it is not enough to match two texts on the purely lexical level, but rather to also take into account the semantic level (see Fig. 1).



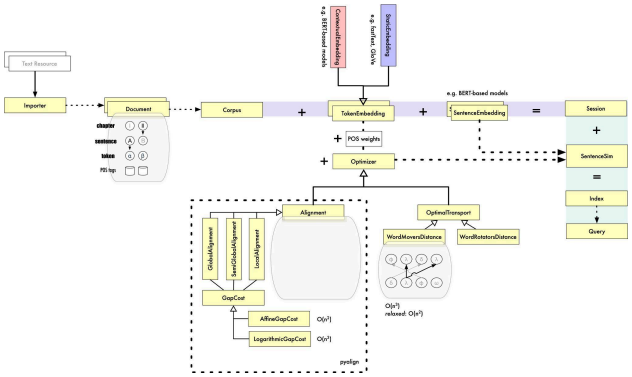
Example for an intertextual reference to Shakespeare's "Macbeth".

The task at hand is typically referred to as *semantic textual similarity* (STS; Cer et al., 2017) and neural text embeddings have long been recognized as a foundational building block of SOTA solutions (Zhelezniak et al., 2020). In recent years, many different approaches to neural embeddings have emerged and have been deemed suitable for different application scenarios. One line of approaches

focuses on providing a single embedding for a span of text, as for example in the case of sentence embeddings (Wang et al., 2021). Another line of approaches uses high quality word embeddings and builds algorithms on top of them, to operate on spans of tokens. Common approaches on this group are optimal transport (Kusner et al., 2015), fuzzy sets (Zhelezniak et al., 2019), or various statistical approaches (Zhelezniak et al., 2020). Interestingly, most of the existing tools for the detection of intertextuality – for instance *Tesserae* (Scheirer et al., 2016) or *Tracer* (Büchler et al., 2014) – do not utilize such neural embeddings at all.

To close this gap, we present the *Vectorian* as an intertextuality search engine (see Manjavacas et al., 2019) that aims to serve as a research framework for running STS queries using established embedding methods on both token and span levels. Besides various types of embeddings, the Vectorian can also combine custom alignment algorithms and further NLP operations, such as the weighting of POS. Two notable features of the Vectorian framework are its fast instantiation of new search indices on pre-processed corpora – including full support for pre-computed static and contextual embeddings – and a fast and optimized alignment search implementation that scales reasonably well to moderately sized corpora.

In our poster, we present the Vectorian API as a software demonstration. The Vectorian can be used to experiment with various configurations of embeddings and alignments for different tasks of intertextuality detection. Figure 2 shows the overall workflow of the Vectorian API [1].



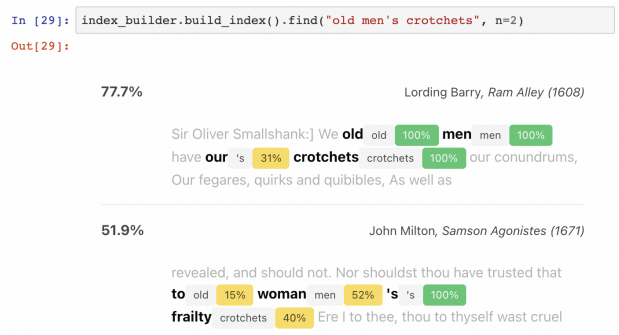
The Vectorian workflow and core elements of the API.

First, the **Importer** is used to preprocess text resources. Essentially, the documents are segmented into tokens and sentences. If the document contains additional structural XML markup, the importer can also be customized to parse this information. Moreover, POS tags are annotated utilizing spaCy. The result of this step is a **Corpus** of segmented and annotated **Documents**. In the next step, the corpus is enriched with contextual information for each word to provide an additional layer for semantic analyses. This is solved via embeddings. At this point, different **TokenEmbeddings** are calculated and stored as vectors. The Vectorian implements various static (e.g. fastText, GloVe) as well as contextual (e.g. BERT-based models) token embeddings.

For technical reasons, **SentenceEmbeddings** are generated in a later step if required. At this stage, all the necessary steps have been taken to instantiate a **Session**, which is an optimized in-memory representation of the given corpus and the selected embeddings. The purpose of a session is to generate a searchable **Index** of the embedded corpus.

For the similarity comparison in **SentenceSim**, first of all a similarity measure (e.g. cosine similarity) is defined. Next, the approach for the actual string comparison is chosen. This can be a local, global, or semi-global **Alignment** approach (Aluru, 2005) with variable gap costs, or the **Word Mover's Distance** (Kusner et al., 2015). Finally, there is an option to use the previously annotated POS as additional weights. The idea of **POS weights** is based on Batanović & Bojic (2015) and ensures that differing tokens that still have the same POS are classified as more similar than if they have a POS mismatch. SentenceSim also allows for an entirely alternative approach to compare the query and document partitions, which is an approach that utilizes the aforementioned **SentenceEmbeddings**. With this approach, sentences are represented as embedding vectors of their own, which means similarity can be directly assessed by comparing sentence vectors.

Once the index has been created, a **Query** can be searched in the previously created corpus. Figure 3 shows an example query for the Shakespeare phrase "old men's crotchets". Two example results that were retrieved by the Vectorian are also provided. These results illustrate how the Vectorian evaluates every word according to the selected embeddings and then provides a score for its match to the original query.



Example results for a query that is matched with the predefined corpus through the Vectorian API.

With our poster, we hope to spark some discussion with the DH community on how to apply and further develop the Vectorian API, which we believe will be a useful resource for any kind of intertextuality research in the DH.

Bibliography

Aluru, S. (Ed.). (2005). Handbook of Computational Molecular Biology. Chapman and Hall/CRC. https://doi.org/10.1201/9781420036275)

Bamman, D. & Crane, G. (2008). The logic and discovery of textual allusion. In In Proceedings of the 2008 LREC Workshop on Language Technology for Cultural Heritage Data.

Batanovic´ & Bojic, 2015). Using Part-of-Speech Tags as Deep Syntax Indicators in Determining Short Text Semantic Similarity. Computer Science and Information Systems, 12(1):1–31, January.)

Büchler, M., Burns, P. R., Müller, M., Franzini, E., & Franzini, G. (2014). Towards a historical text re-use detection. In Text Mining (pp. 221-238). Springer, Cham.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). SemEval-2017 Task 1: Semantic Textual Similarity - Multilingual and Cross-lingual Focused Evaluation. https://doi.org/10.18653/v1/S17-2001

Kusner, M. J., Sun, Y., Kolkin, N. I., & Weinberger, K. Q. (n.d.). From Word Embeddings To Document Distances.

Manjavacas, E., Long, B., & Kestemont, M. (2019). On the Feasibility of Automated Detection of Allusive Text Reuse. arXiv:1905.02973 [cs], May.

Scheirer, W., Forstall, C., & Coffee, N. (2016). The sense of a connection: Automatic tracing of intertextuality by meaning. Digital Scholarship in the Humanities, 31(1), 204-21.

Zhelezniak, V., Savkov, A., Shen, A., Moramarco, F., Flann, J., Hammerla, N. Y., & Health, B. (2019). Don't settle for average, go for the max: Fuzzy sets and max-pooled word vectors.

Zhelezniak, V., Savkov, A., Hammerla, N., & Health, B. (2020). Estimating Mutual Information Between Dense Word Embeddings.

Wang, K., Reimers, N., & Gurevych, I. (2021). TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning.

Notes

1. Vectorian API: https://poke1024.github.io/vectorian/index.html

A Unicode Input Support Tool for Searching Chinese Characters by Components and Stroke Number

Liu, Guanwei
k-ryu@hi.u-tokyo.ac.jp
University of Tokyo, Japan

Nakamura, Satoru
nakamura@hi.u-tokyo.ac.jp
University of Tokyo, Japan

Yamada, Taizo
t_yamada@hi.u-tokyo.ac.jp
University of Tokyo, Japan

In recent years, given that the Unicode Standard has been updated, the number of Chinese character (Hanzi/Kanji) codes available for text databases of historical documents written in Chinese and Japanese has increased significantly. With the addition of the CJK Unified Ideographs Extension G in 2020, the current Unicode Standard now contains a total of more than 93,000 Chinese characters (Lunde, 2021). To maximize reproduction of the content in historical documents, it is desirable to use Chinese characters that can be displayed in a form that is as close to the original glyphs as possible.

TEI-encoded [1] text databases are also becoming increasingly common. However, there are still few examples of the use of these extended Chinese characters in publicly available text databases. Possible causes are the difficulty of checking whether a character can be represented in Unicode and determining how to input those newly included characters in Unicode without input method support.

The most common computer-based Chinese character input method entails reading the character. If the reading is inconclusive or if the character has not yet been incorporated into an input method, Ideographic Description Sequence (IDS) can be used to search for the input (The Unicode Consortium, 2021).

IDS describes Chinese characters according to their mechanism and components. For example, 紅(red) can be described as ⿰糸工 (Figure 1). Tools such as CHISE [2] (Morioka, 2008) and GlyphWiki [3] (上地, 2018) are useful to search for Chinese characters (in Japanese), but no existing tools specialize in inputting Chinese characters and allow TEI-encoded output data in a format suitable for reprinting.

Regarding actual reprinting, there are three issues that need to be addressed.

1. Searching: Transcribing old documents is complex on a computer. Moreover, it is difficult to input components as search keywords (e.g., 氵 or 羊).
2. Display: Results display depends on the font installed on the device; characters that are not part of the font will be displayed in Tofu. Without a suitably designed results page, it is difficult to select and copy Chinese characters for input.
3. Application: Searching should be fast and accurate. Given that it is necessary to search for all Chinese characters in Unicode, deal those data may takes a lot of time before the results are displayed.

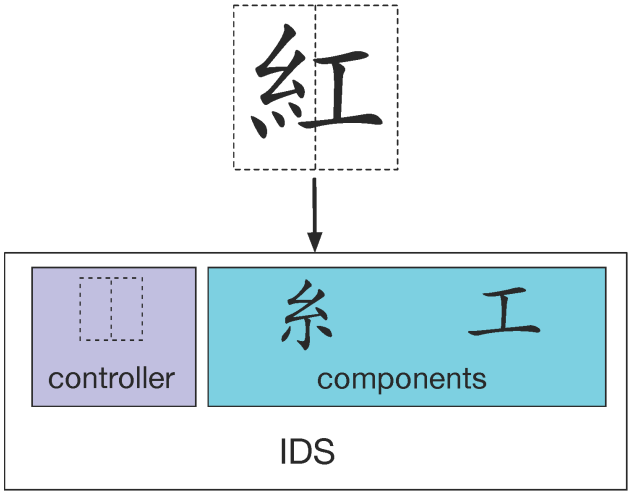


Figure 1:
a sample of IDS

To solve the above issues, we have developed a tool for searching and inputting Chinese characters that cannot be input via ordinary input methods using Chinese character components and stroke numbers. The tool has the following features, and the corresponding issue is given in parentheses:

1. With data on characters' components from CHISE. IDS [4], can search up to 93,000 characters based on the newest Unicode Standard. [Searching]
2. Filter results by the number of strokes remaining. The data stroke number is from Unihan [5]. [Searching]
3. Decomposition of Chinese characters to input complex parts. [Searching]
4. Search results are displayed as SVG or PNG images from GlyphWiki to avoid displaying in Tofu.
5. Copy the target character to the clipboard with one click. [Display]
6. The user does not need to have fonts to cover all the Chinese characters in Unicode because the results display as SVG glyphs from GlyphWiki. [Display]
7. Data are cached in the memory to avoid communication latency with a server. [Application]
8. Output in multiple formats (encoded character, Unicode scala, XML block). XML block is customizable with template (Figure 3). [Application]

The tool uses IDS data from CHISE and glyph images from GlyphWiki. Compared to both, users can search and input Chinese characters more efficiently by filtering according to the number of strokes remaining (Figure 2). Furthermore, it is easy to select and copy the result in one click, offering a fast response. It is also possible to copy the Chinese characters you want to input as a block of TEI-encoded XML to create a text database. This saves database creators' time. Here is an example (Figure 4).

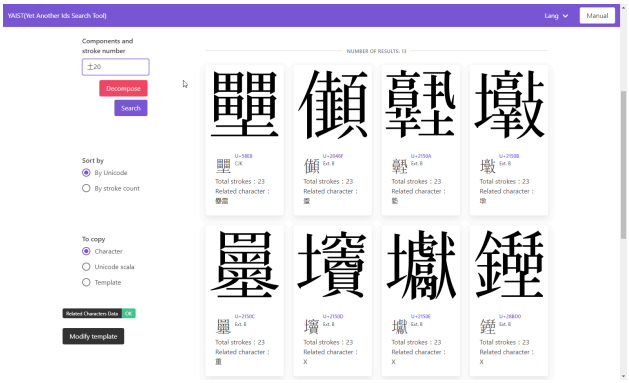


Figure 2:
interface of the tool

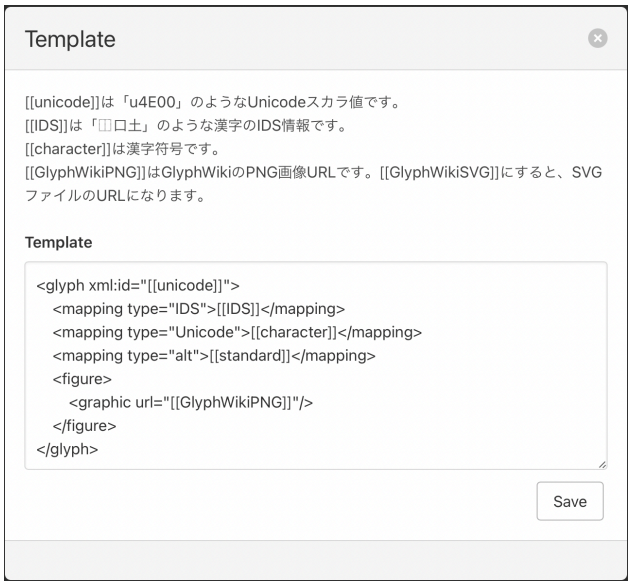


Figure 3:
customizable XML block template

```
<glyph xml:id="u3ECE">
  <mapping type="IDS">⿰大⿱王王</mapping>
  <mapping type="Unicode">爽</mapping>
  <mapping type="standard">瑟</mapping>
  <figure>
     <graphic url="https://glyphwiki.org/glyph/u3ece.png"/>
  </figure>
</glyph>
```

Figure 4:
TEI-encoded sample

This tool has been applied to the creation of TEI-encoded text databases at the Historiographical Institute, University of Tokyo [6]. Details will be presented in the poster.

Bibliography

Lunde, J. H. J. C. K. (2021). *Unicode® Standard Annex #38 UNICODE HAN DATABASE (UNIHAN)* https://www.unicode.org/reports/tr38/#BlockListing, (accessed 20 November 2021).

Morioka, T. (2008). *CHISE: Character Processing Based on Character Ontology*. In Tokunaga, T. and Ortega, A. (eds), Large-Scale Knowledge Resources. Construction and Application. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 148–62 https://doi.org/10.1007/978-3-540-78159-2_14.

The Unicode Consortium (2021). *The Unicode® Standard Version 14.0*. https://www.unicode.org/versions/Unicode14.0.0/ch18.pdf, (accessed 20 November 2021).

上地, 宏一(2018). *GlyphWiki*：漢字字形自由共有サイト. 学芸国語国文学, 48: 181–91 doi:10.24672/gkokugokokubun.48.0_181.

Notes

1. https://tei-c.org/
2. https://www.chise.org/
3. https://www.glyphwiki.org/
4. https://gitlab.chise.org/CHISE/ids
5. https://www.unicode.org/Public/UCD/latest/ucd/Unihan.zip
6. https://www.hi.u-tokyo.ac.jp/en/

Centering the Marginalized: Scholar-Curated Worksets from the HathiTrust Digital Library

Magni, Isabella
magni.isabella@gmail.com
HathiTrust Research Center, United States of America

Worthey, Glen C.
gworthey@illinois.edu
HathiTrust Research Center, United States of America

Graham, Maryemma
mgraham@ku.edu
HathiTrust Research Center, United States of America

Walsh, John A.
jawalsh@indiana.edu
HathiTrust Research Center, United States of America

Downie, J. Stephen
jdownie@illinois.edu
HathiTrust Research Center, United States of America

Dubnicek, Ryan C.
rdubnic2@illinois.edu
HathiTrust Research Center, United States of America

The *Scholar-Curated Worksets for Analysis, Reuse & Dissemination* (SCWAReD) project, generously supported by the Andrew W.Mellon Foundation, is producing a suite

of scholar-curated, targeted worksets of materials from the HathiTrust Digital Library, facilitated by its Research Center (HTRC). HTRC worksets are user-created collections of HathiTrust volumes that can be treated as data and analyzed using a variety of tools. Worksets can be shared and cited, contributing to research reproducibility and durable scholarship. In addition to their value as focused digital collections, SCWAReD's scholar-curated worksets also serve as illustrative, *reusable research models*, and include not only the worksets themselves, but also scholarly introductions, derived datasets and related documentation, and research reports, demonstrating the collaborative workset-building, textual analysis, workflow development, and dataset creation activities typically carried out by HTRC.

The special mission of SCWAReD is to highlight and center the work of historically under-resourced and marginalized textual communities. For this purpose, a flagship project and four sub-projects were selected competitively; each of them explores new methods for creating, analyzing, and reusing curated digital collections and the research data derived from them. The need to address inequities in both library collections and digital humanities research is already well documented (e.g. among others: Gallon, 2016; McPherson, 2012; Earhart, 2012). SCWAReD aims to help address these inequities in both library collections and digital research by identifying and remediating gaps within HathiTrust, and by using computationally-assisted efforts to recover content that is already part of the HathiTrust Digital Library but may be difficult to discover with traditional metadata, in a traditional catalog, from within a massive digital collection.

SCWAReD's flagship collaboration is with the Black Books Interactive Project, part of the longstanding History of Black Writing (HBW), founded in 1983 at the University of Mississippi by SCWAReD Co-PI Maryemma Graham and hosted since 1998 under her leadership at the University of Kansas.

Four more projects were selected to create curated worksets to be developed concurrently:

- "Mining the Native American Authored Works in HathiTrust for Insights," in which directors Kun Lu, Raina Heaton, and Raymond Orr (University of Oklahoma) seek to develop a database of Native American authors and their bibliographic information, create a reusable workset of Native American authored works in HathiTrust, and provide insights into the characteristics of the community by text mining their works;
- "The Black Fantastic: Curated Vocabularies, Artifact Analysis and Identification," in which directors Clarissa West-White (Bethune Cookman University) and Seretha Williams (Augusta University) propose to prove that characteristics of the Black Fantastic—the cultural production of African Diasporic artists and creators who engage with the intersections of race and technology in their work—exist in historical and current cultural artifacts, including those created by and about future-forward personalities, such as Dr.Mary McLeod Bethune;
- "Creating Period-Specific Worksets for Latin American Fiction," in which director José Eduardo González (University of Nebraska, Lincoln) seeks to create datasets to research the history of Latin American fiction and question traditional periodization of this literature by attempting to detect the boundaries between literary periods and subgenre distinctions; and
- "The National Negro Health Digital Project: Recovering and Restoring a Black Public Health Corpus," in which director Kim Gallon (Purdue University) draws on HathiTrust's collection of public health documents on Black health to explore how early twentieth century Black public health officials communicated and addressed health disparities that impacted African American communities.

For each of these projects, we identify and attempt to fill collection gaps (items documented by scholar-curators, but missing from HathiTrust). We also create, collect, and document our research artifacts (elements of a "reusable research model," as described above), and include them with the curated workset. These include the search algorithms devised for the survey of existing holdings; data derived from the workset objects; bibliographies and bibliographic essays; curatorial statements; and whatever other apparatus and artifacts may be deemed significant for interpreting and analyzing the workset, or amenable for later reuse, all of which will be released open access. In each of these partnerships, project teams bring content and domain expertise, research questions, and curation experience, while HTRC provides HathiTrust collection access, research tools and environments, and technical expertise in text and data mining. Research questions suitable for interrogation in HathiTrust holdings have been developed in the course of each project, informed by the workset building process, available content, and gaps identified.

Our poster will provide an overview of the SCWAReD project, our flagship collaboration with the Black Books Interactive Project, and our four collaborative projects. We will also provide preliminary results and report on gap-filling efforts.

Bibliography

Earhart, A. (2012). Can Information Be Unfettered? Race and the New Digital Humanities Canon. In Gold, M. ed. *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press, chapter 18, https://doi.org/10.5749/9781452963754 (accessed 27 April 2022).

Gallon, K. (2016). Making a Case for the Black Digital Humanities. In Matthew Gold, ed. *Debates in the Digital Humanities 2016*. Minneapolis: University of Minnesota Press, chapter 4, https://doi.org/10.5749/9781452963761 (accessed 27 April 2022).

McPherson, T. (2012). Why Are the Digital Humanities So White? In Gold, M. ed. *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press, chapter 9, https://doi.org/10.5749/9781452963754 (accessed 27 April 2022).

Responding to Tibetan Diversity: Rebuilding the Mandala Scholarly Content Management System

Mapp, Rennie
mapp@virginia.edu
U of Virginia, United States of America

Shinozaki, Yuji
ys2n@virginia.edu
U of Virginia, United States of America

Gunn, Stan
stg2s@virginia.edu
U of Virginia, United States of America

Introduction: Mandala is a longstanding Digital Humanities project at the University of Virginia that contains significant Tibetan-Himalayan cultural heritage materials, as well as custom software tools and an ontology framework. As of 2021, a UVA Library development team is undertaking the restructuring and updating of the Mandala ontology framework, focusing on bringing Mandala's original metadata schema into alignment with other open-access schemata for cultural heritage materials through the adoption of linked data and the creation of an ontology-builder user interface (UI). Our team's poster presents a roadmap of the ontology redevelopment process, identifies particular challenges associated with the current "Mandala Knowledge Maps" ontology and the Tibetan-Himalayan materials it describes, and offers an overview of the planned UI, which we hope to make generally useful for digital cultural heritage research.

Background: Mandala is one of the world's richest and most diverse digital collections of Tibetan and Himalayan culture, with resources for the general public, literary scholars, anthropologists, archeologists, and other researchers and educators. Over the last three decades, under the leadership of Professor David Germano, Mandala has been documenting this culture with textual editions, translations, dictionaries, and encyclopedias, as well as audio-video recordings of oral traditions, music, histories, philosophies, environmental knowledge, and more. Mandala also stores many thousands of photographic images of cultural activities, architecture, art, and the physical environment, and ontologies of various cultural subjects, as well as extensive data on geographical features. Mandala's descriptive ontology, called "Knowledge Maps" (or Mandala KMaps), was first developed in the early 2000s. Professor Germano and his colleague in the Tibet Center, Andres Montano, developed ontologies to describe geographic locations over time, topic maps (subjects), and Tibetan terms. This sophisticated and rich ad hoc ontology has been developed over decades. However, the current architecture does not employ open standards, and thus remains somewhat limited in its openness and reusability.

Project Goals: Mandala's cultural heritage assets and the framework that supports them are now sponsored by the University of Virginia Library's Information Technology Department, which is undertaking to ensure that these digital assets are preserved in a sustainable way. The Mandala 2.0 team is developing a readily accessible and reproducible descriptive framework so that these valuable resources are available more broadly, using standardized technologies: under consideration are RDF, Wikidata and Linked Open Data more generally. The team also plans to create a user interface to enable researchers new to ontology-building to create and manage new metadata schemata describing their research in the humanities and humanistic social sciences. The Mandala 2.0 Project is managed by Rennie Mapp, with Yuji Shinozaki as Technical Director and Stan Gunn as Executive Director of UVA Library IT.

Poster Contents: The poster will include a structural diagram of the current Mandala technology stack; a diagram of the proposed structure for Mandala 2.0; examples of the current KMaps ontology structures; a table of standard, open technology frameworks for ontology building under consideration, along with advantages and disadvantages for each; a timeline for development; and short narratives to describe the project's past and proposed future.

Bibliography

Cimiano, Philipp, et al. (2020). *Linguistic Linked Data Representation, Generation and Applications*. Springer International Publishing.

Cobb, Joan (2015). "The Journey to Linked Open Data: The Getty Vocabularies." Journal of Library Metadata, vol. 15, no. 3/4, July 2015, pp. 142–156. EBSCOhost, doi:10.1080/19386389.2015.1103081.

Chanhom, Weeraphan, and Chutiporn Anutariya (2019). "TOMS: A Linked Open Data System for Collaboration and Distribution of Cultural Heritage Artifact Collections of National Museums in Thailand." *New Generation Computing*, vol. 37, no. 4, Dec. 2019, pp. 479–498. EBSCOhost, doi:10.1007/s00354-019-00063-1.

Nurmikko-Fuller, Terhi (2018). "Publishing Sumerian Literature on the Semantic Web." *CyberResearch on the Ancient Near East and Neighboring Regions: Case Studies on Archaeological Data, Objects, Texts, and Digital Archiving*, edited by Vanessa Bigot Juloux et al., Brill, 2018, pp. 336–64, http://www.jstor.org/stable/10.1163/j.ctv4v349g.20.

Robledano-Arillo, Jesús, et al (2020). "Application of Linked Open Data to the Coding and Dissemination of Spanish Civil War Photographic Archives." *Journal of Documentation*, vol. 76, no. 1, Jan. 2020, pp. 67–95. EBSCOhost, doi:10.1108/JD-06-2019-0112.

Wikisource contributors (2019). "ARL White Paper on Wikidata Opportunities and Recommendations." Wikisource . Wikisource , 15 May. 2019. Web. 7 Dec. 2021.

Extraction and Automatic Generation of Characters' Attributes in Contemporary Japanese Entertainment Works

Murai, Hajime
h_murai@fun.ac.jp
Future University Hakodate, Japan

Toyosawa, Shuuhei
g2120029@fun.ac.jp
Future University Hakodate, Japan

Shiratori, Takayuki
g2120018@fun.ac.jp
Future University Hakodate, Japan

Yoshida, Takumi
g2120050@fun.ac.jp
Future University Hakodate, Japan

Nakamura, Shougo
g2121042@fun.ac.jp
Future University Hakodate, Japan

Saito, Yuuri
g2121021@fun.ac.jp
Future University Hakodate, Japan

Ishikawa, Kazuki
g2121004@fun.ac.jp
Future University Hakodate, Japan

Nemoto, Sakura
g2121045@fun.ac.jp
Future University Hakodate, Japan

Iwasaki, Junya
g2121007@fun.ac.jp
Future University Hakodate, Japan

Ohta, Shoki
b1018131@fun.ac.jp
Future University Hakodate, Japan

Ohba, Arisa
b1018174@fun.ac.jp
Future University Hakodate, Japan

Fukumoto, Takaki
b1018230@fun.ac.jp
Future University Hakodate, Japan

Introduction

It has been long established by several studies that it is possible to extract the common plot structure of specific genre stories when many such stories are collected (Barthes 1968, Propp 1968, Campbell 1949). Based on these old humanistic studies, recent research focusing on several specific genre stories has clarified that the quantitative and objective extraction of common plot structures can be executed using computational methods (Murai 2014, 2020).

In these recent studies, the plot structures were described as sequences of symbolized scenes or functions. The common plot structures of specific genres were extracted using quantitative methods for the symbolized sequences.

However, these past studies focused only on the plot structures and therefore, the quantitative features of the characters in the stories have not been focused. The present study is the first to develop a common symbol set to describe the character's attributes as well as the plot structures of several different genres. The identification of symbols that are common between different story genres enables a comparison of the characteristics of each story genre. These symbol sets can be used to extract the common patterns of general stories. Moreover, the extracted common patterns could become the foundation for automatic story generation systems.

Target contents

To compare different story genres, several popular genres in modern Japanese entertainment culture were selected based on the several famous comic and game sales rankings (Murai 2021). The selected genres were "Adventure," "Battle," "Love," "Detective," and "Horror."To extract typical plot structures for each genre, works of combined genres (such as "love comedy") were eliminated, and popular short stories were selected based on sales rankings. If there were too few popular short stories, popular long stories were divided into short stories based on changes in the purpose of the stories' protagonists (Nakamura 2020). Subsequently, the selected stories were divided into plot elements (scenes) which were categorized based on a common hierarchical category (Murai 2021).

Development of automatic character generation program

	Stories	Characters	Average characters
Adventure	206	1256	6.1
Battle	243	2245	9.2
Love	123	541	4.4
Detective	134	929	6.9
Horror	167	585	3.5
Total	873	5556	6.4

Table 1.
Stories and characters for each genre

Approximately 5,500 characters were extracted from about 900 stories in five genres. These characters were digitized based on the character attribute category that included six areas: gender, age, blood relationship, role in the story, social position, and species. Each area included several attributes and the total number of attributes was 43. The characters' features were investigated and appropriate attributes were assigned. This process was based on discussions among several analysts of narratology.

One character was often assigned several attributes, for instance, "male," "young," "elder brother," "enemy," and "soldier."

In the next step, a chi square test was performed for the assigned attributes of the extracted characters, and frequently and rarely appearing attributes for each genre were identified. Moreover, a co-occurrence analysis was performed, and frequently co-occurring pairs of attributes were extracted.

Based on the analyzed statistical features, an automatic story character generation program was developed, which output an appropriate number of characters and a combination of attributes for each character when a genre was input. These outputs were based on frequently appearing attributes and frequently co-occurring pairs of attributes in each genre.

In the automatic story character generation program, a typical pattern of characters in existing stories could be generated. In addition, by adjusting the parameters of the probability of appearance, rare patterns that seldom occurred in existing stories could also be generated. Therefore, this program can be applied not only for genre-dependent typical works, but also for works with high novelty.

Conclusion

A data set of attributes for story characters was developed by utilizing a cross-genre story data set. In addition, based on a statistical analysis of genre-dependent attributes of story characters, an automatic story character generation program was developed. This program will be applied in an automatic story generation system along with an automatic plot generation program (Murai 2021). In future, the complete automatic story generation system will output stories according to the user-selected genre.

Bibliography

Barthes, R. (1968). Elements of Semiology. Hill and Wang, New York, USA.

Campbell, J. (1949). The Hero with a Thousand Faces. Pantheon Books, USA.

Murai, H. (2014). "Plot Analysis for Describing Punch Line Functions in Shinichi Hoshi's Microfiction", 2014 Workshop on Computational Models of Narrative, (Eds.

Mark A. Finlayson, Jan Christoph Meister, and Emile G. Bruneau), OpenAccess Series in Informatics, 41:121-129.

Murai, H. (2020). “Factors of the Detective Story and the Extraction of Plot Patterns Based on Japanese Detective Comics”, Journal of the Japanese Association for Digital Humanities, 5(1): 4-21.

Murai, H., Toyosawa S., Shiratori, T., Yoshida, T., Nakamura, S., Saito, Y., Ishikawa, K., Nemoto, S., Iwasaki, J., Uda, A., Ohta, S., Ohba, A., and Fukumoto, T. (2021). “Dataset Construction for Cross-genre Plot Structure Extraction”, Proceedings of the Japanese Association for Digital Humanities, 93-96.

Nakamura, S. and Murai, H. (2020). “Proposal of a Method for Analyzing Story Structure of Role-playing Games Focusing on Quest Structure”, Computer and Humanities Symposium, 2020: 149-156 (in Japanese).

Propp, V. (1968). Morphology of the Folk Tale. U of Texas P, USA.

Building a Knowledge Base for Data-Driven Historical Information Research Infrastructure and Its Application with Historical Painting Materials

Nakamura, Satoru
na.kamura.1263@gmail.com
The University of Tokyo

Suda, Makiko
suda@hi.u-tokyo.ac.jp
The University of Tokyo

Kuroshima, Satoru
kurosima@hi.u-tokyo.ac.jp
The University of Tokyo

Inoue, Satoshi
inoue@hi.u-tokyo.ac.jp
The University of Tokyo

Yamada, Taizo
t_yamada@hi.u-tokyo.ac.jp
The University of Tokyo

Introduction

The Historiographical Institute in the University of Tokyo has provided a digital archive called “SHIPS” with about 40 different databases (a total of 5.6 million data) and about 20 million images of historical materials related to Japan from the 8th to the 19th century. Based on these accumulated data, we aim to establish a system that supports historical data analysis and visualization by applying information technology to realize the advancement and efficiency of the historical research process and the multifaceted dissemination of research results, including educational use. For this purpose, we build a knowledge base about people, places, and historical facts, and create an application that utilizes this knowledge base. This is an effort to elucidate the effects of data-driven methods on the research process of history. This study targets 2 historical painting materials; Wako Zukan and Shoho Ryukyu Kuniezu, owned by the Historiographical Institute.

The construction of a data-driven infrastructure for historical information research is a theme actively pursued both in Japan and abroad. In Europe, the Time Machine Project [3] is underway, and in Japan, for example, ROIS/ Collaborative Open Data Center for the Humanities is conducting historical big data research [1]. We will also carry out research using international standards such as TEI, IIIF, and RDF to build a highly interoperable research infrastructure that can be linked to these research results. It will enable the utilization of research resources across institutions and countries.

Case 1: Wako Zukan

The Wako Zukan is a historical document depicting the appearance of Japanese pirates in the 16th century. This is a picture scroll over five meters long and depicts the story of the Ming dynasty's defeat of the Japanese pirates who attacked from across the sea, but since there are no characters in the picture scroll to determine the character of the scroll, reading it has been a major challenge.

We have developed a digital storytelling feature that utilizes IIIF annotations to solve this problem. The annotations are displayed along with the screen transitions of the scroll, and the image portions are magnified. The IIIF's Choice function also provides a comparison screen with infrared photographs. These functions enable us to express the contents of the picture scroll interactively.

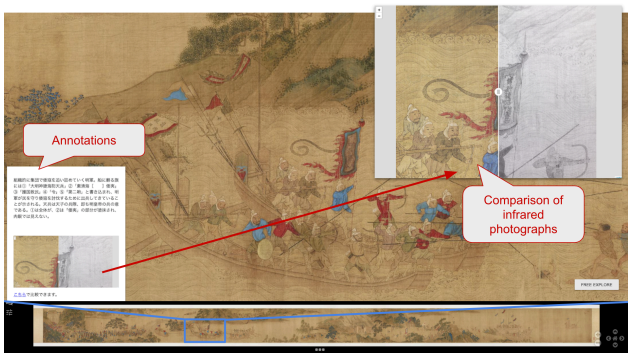


Fig. 1
Digital storytelling using IIIF annotations and infrared photographs.

We have also developed a text viewer using IIIF and TEI. It displays several types of text on the left, images in the center, and maps and lists of people and places on the right. These data are structured by TEI and displayed with each other. This supplementary information supports the reading of the text.



Fig. 2
Text viewer that displays several types of text, images, and maps in conjunction.

These features have enabled us to incorporate the knowledge that historians have been deciphering.

Case 2: Shoho Ryukyu Kuniezu

The Shoho Ryukyu Kuniezu are large-scale maps with a maximum length of over 7 meters and were produced nationwide in 1644 under the orders of the Edo Shogunate. It is the oldest large-scale pictorial map of the Ryukyu Islands and has attracted much attention because of its rich topographical and textual information, but the problem was that it was difficult to handle due to its size.

We developed a system that provides functions including viewing high-resolution images and text searches of written information to solve this problem. Annotations were assigned to each place name, giving information on classification and latitude/longitude.

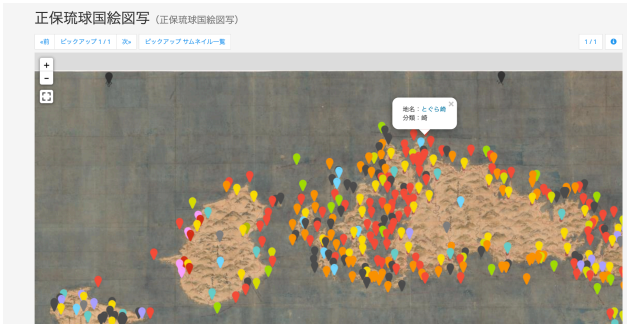


Fig. 3
Annotations and searches for place names written in kuzushiji.

It also provides a function to superimpose the sketch map with a modern map based on the coordinate information given to the annotations (Fig. 4 Left). In addition, a comparison of kuniezu from different periods released by the National Archives of Japan revealed differences in the precision of topographical rendering (Fig. 4 Right).

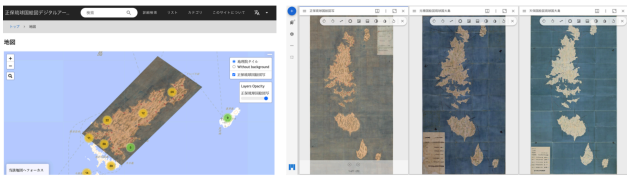


Fig. 4
Comparison of images with modern maps and materials from other institutions.

Structuring the data with annotations enables us to solve the problem of reading kuzushiji, which is a high hurdle for many users. In addition, metadata for place names enables the data to be overlaid on modern maps, making it possible to expand the use of the data to include research other than historical studies.

Conclusion

We describe the construction of a knowledge base for building a data-driven infrastructure for historical information research. We have published these applications [2][4] in December 2021. These applications' actual and potential use will be reported at the conference.

Bibliography

Asanobu KITAMOTO and Mika ICHINO and Chikahiko SUZUKI and Tarin CLANUWAT. (2018). *Historical Big Data: Reconstructing the Past through the Integrated Analysis of Historical Data. Eighth Conference of Japanese Association for Digital Humanities (JADH2018),* pp.67-69.

Shoho Ryukyu Kuniezu Digital Archive. https://www.hi.u-tokyo.ac.jp/collection/degitalgallary/ryukyu/en/ (accessed 10 December 2021).

Time Machine Europe. https://www.timemachine.eu/ (accessed 10 December 2021).

Wakozukan Digital Archive. https://www.hi.u-tokyo.ac.jp/collection/degitalgallary/wakozukan/en/ (accessed 10 December 2021).

Introducing MPCD – Middle Persian Corpus and Dictionary

Neuefeind, Claes
c.neuefeind@uni-koeln.de
University of Cologne, Germany

Mondaca, Francisco
f.mondaca@uni-koeln.de
University of Cologne, Germany

Eide, Øyvind
oeide@uni-koeln.de
University of Cologne, Germany

Colditz, Iris
iris.colditz@rub.de
Ruhr-University Bochum, Germany

Jügel, Thomas
thomas.juegel@rub.de
Ruhr-University Bochum, Germany

Rezania, Kianoosh
kianoosh.rezania@rub.de
Ruhr-University Bochum, Germany

Cantera, Alberto
Alberto.Cantera@fu-berlin.de
Free University Berlin, Germany

Emanuel, Chagai
chagai17@gmail.com
Hebrew University Jerusalem, Israel

Introduction

The project "Zoroastrian Middle Persian – Digital Corpus and Dictionary (MPCD)"[1] aims at creating a comprehensive, open-access corpus of Zoroastrian Middle Persian texts in Pahlavi script, accompanied by a digital Middle Persian-English dictionary based on this corpus. Started in mid 2021, MPCD is funded by the DFG as a long-term project[2], with a duration of nine years in total. The cooperative project is being carried out at the universities of Bochum, Berlin, Jerusalem and Cologne.

While the partners in Bochum, Berlin and Jerusalem focus on the philological aspects of the project, the Cologne Center for eHumanities (CCeH) is responsible for the technical implementation of a collaborative working environment, which at the same time serves as user interface for research and analysis of the processed resources. Of key importance to both the philological work and the technical design of the application is a common data model, which thus will be addressed in this poster.

Scope of the project

Middle Persian was the official language and lingua franca of the Sasanian Empire (3rd-7th century) and was of high cultural and supra-religious importance. From late antiquity to the early Islamic period it connected the different areas of the Iranian East and West in both linguistic and cultural terms. However, the extensive corpus of Middle Persian texts has only been partially indexed to date and there is no comprehensive lexicographical resource covering the full variety of its vocabulary.

The aim of the MPCD project is to fill this gap by creating a corpus of all Zoroastrian Middle Persian texts in Pahlavi script (about 54 texts, approx. 687.000 words). This corpus will be made accessible in transliteration and transcription (cf. Rezania 2020) as well as in manuscript photographs of the 15 oldest codices, some of which can be obtained from Alberto Cantera's CAB project[3] (Corpus Avesticum Berolinense). This comprehensive digital corpus will subsequently be used as a basis for the creation of a digital Middle Persian-English dictionary, expected to comprise approx. 7.000 lemmata.

With its close interlocking of text and dictionary, the project complements existing text collections on Middle Persian such as TITUS (Thesaurus of Indo-European Text and Language Materials)[4] and extends existing concise dictionaries such as MacKenzie (1971) or Nyberg (1964/1974). The project is conceived as a basis for identifying internal and external factors in the complex fabric of the texts of Zoroastrian Middle Persian literature, and for providing an adequate means for a differentiated analysis of cultural, religious and social history.

Modeling MPCD

The digital corpus and dictionary represent two closely interlocked analytical tools with different emphases – text structure and semantics – that are also closely intertwined in the work process. This has to be taken into account by the internal data model, which at the same time determines the corpus structure and the collaborative working environment.

At the current stage of the project we focus on the corpus model. The corpus consists of texts (element <corpusText> in figure 1), with each text holding basic metadata and a number of sentences. The metadata comprises basic information like sigle, title, creation date, source, the responsible editor and his/her collaborators, while each <sentence>-element contains the full sentence, one or more translations, an (optional) comment and finally one or more tokens.

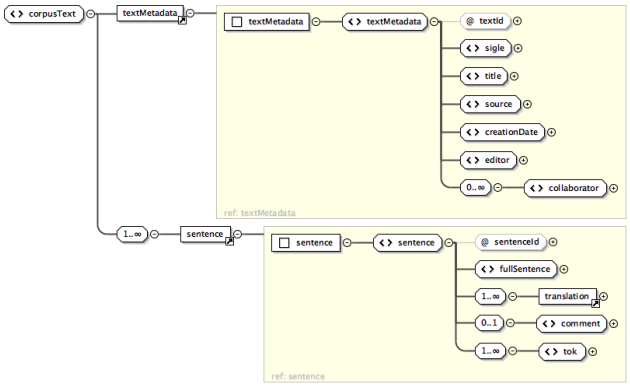


Fig. 1:
Excerpt of the corpus model reflecting a single text (element <corpusText>).

Each token (element <tok> in figure 2) holds information on the token language, a transcription and a transliteration, its lemma, (optional) information on the text structure to mark the beginning of a new section, folio etc (element <newPart>). Besides that, the token model includes both morphosyntactic annotations and lexicographic information, where the latter will prospectively serve as a direct link to the corpus-based dictionary (see element <dictRef> in <tokenSemantics>).

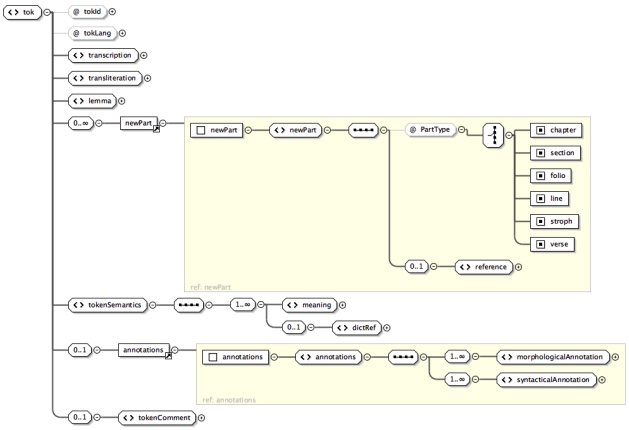


Fig. 2:
Excerpt of the corpus model reflecting a single token (element <tok>).

The morphosyntactic annotations will largely follow the *Universal Dependencies* [5] standard, which is adapted for the MPCD project by determining the subset of tags necessary for the annotation of Middle Persian and by adding Pahlavi-specific tags. These fine-grained linguistic annotations on token-level will allow for differentiated searches according to linguistic parameters that will be implemented on the basis of elasticsearch; search and CRUD operations will be available via a GraphQL-API (cf. Mondaca et al. 2019a and 2019b).

With its focus on the data model, the poster will provide a compact overview of the MPCD project, reflecting the corpus structure, the transcription process and the philological decisions as well as the implications for the technical design of the working environment to be established.

Bibliography

MacKenzie, D. N. (1971): A Concise Pahlavi Dictionary. London/New York/Toronto.

Mondaca, F., Rau, F., Neuefeind, C., Kiss, B., Kölligan, D., Reinöhl, U., Sahle, P. (2019a): *C-SALT APIs - Connecting and Exposing Heterogeneous Language Resources.* In: Book of Abstracts of the Digital Humanities Conference 2019 (DH2019) 09.07-12.07.2019. Utrecht, Netherlands.

Mondaca, F., Schildkamp, P., Rau, F. (2019b): *Introducing Kosh, a Framework for Creating and Maintaining APIs for Lexical Data.* In: Electronic Lexicography in the 21st Century. Proceedings of the

eLex 2019 Conference, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., pp. 907–921.

Nyberg, H.S. (1964): *A Manual of Pahlavi. Part I: Texts, Alphabets, Index, Paradigms, Notes and an Introduction.* Wiesbaden.

Nyberg, H.S. (1974): *A Manual of Pahlavi. Part II: Ideograms, Glossary, Abbreviations, Index, Grammatical Survey, Corrigenda to Part I.* Wiesbaden.

Rezania, K. (2020): *A Suggestion for the Transliteration of Middle Persian Texts in Zoroastrian Middle Persian: Digital Corpus and Dictionary (MPCD): A Three Layered Transliteration System.* Estudios Iranios y Turanios 4: pp. 153–73.

Notes

1. https://mpcorpus.org/
2. https://gepris.dfg.de/gepris/projekt/452473565?language=en
3. https://www.geschkult.fu-berlin.de/en/e/iranistik/forschung/CAB/index.html
4. https://titus.uni-frankfurt.de/
5. https://universaldependencies.org/

Do we have to limit our research question by the tool to be used? The iLCM as an example of freely extensible research software for text-based research tasks in the humanities

Niekler, Andreas
aniekler@informatik.uni-leipzig.de
Leipzig University, Germany

Kahmann, Christian
kahmann@informatik.uni-leipzig.de
Leipzig University, Germany

The barriers to our creativity are often the tools we have at our fingertips. This is particularly visible in text-oriented research tasks in the humanities, where a wide range of libraries and stand-alone software are available. The constraints of each tool and the researcher's skills in using the tools both shape and co-construct the research process. In our view, such influence should be minimized in the research process. The interactive Leipzig Corpus Miner (iLCM) [https://ilcm.informatik.uni-leipzig.de/] (Niekler et. al., 2018) represents a ready-to-use and GUI-supported software solution for the use of text mining in the humanities, cultural studies, and social sciences. The software is completely based on R [https://www.r-project.org/] and RShiny [https://shiny.rstudio.com/]. In parallel to the iLCM interface, an RStudio server [https://www.rstudio.com/] instance is provided as an IDE that ensures access to the available data and results. Among other things, the tool offers the pre-processing of multilingual documents, the retrieval and management of document collections, the deduplication of content, the analysis of word frequency, the analysis of word co-occurrence, time series analysis, topic models, the automatic coding and annotation of categories, supervised text classification (e.g. sentiment analysis) and more. In our poster, we demonstrate that its built-in ability to produce custom scripts, export results and script-based adaptations of the available analyses circumvents some restrictions of other tools used in humanities research.

For researchers using text-oriented analysis methodologies, implementing their own analysis programs is often not a viable option. While more and more researchers in the humanities, cultural studies, and social sciences are acquiring programming skills, developing complex research software remains a complicated process that typically requires trained software developers. Instead, applied research relies on existing research software. There are many single, specific (i.e. for one purpose) ready-to-use tools, e.g. for topic modeling [e.g. https://dariah-de.github.io/TopicsExplorer/], concordance tools [e.g. https://voyant-tools.org/] or word scaling [e.g. http://www.wordfish.org]. This restricts researchers to narrow study designs that stay within the confines of the specific software which leads to a significant reduction in the method portfolio of a research project. If scholars want to map more complex (NLP) workflows it gets more difficult, because researchers often have to use full-blown frameworks or APIs [https://www.nltk.org/, https://spacy.io/, https://stanfordnlp.github.io/CoreNLP/, https://opennlp.apache.org/, https://quanteda.io/] which is a deterrent for many humanists. Digital humanities tools must therefore not only implement general best practices from the field of usability, but also need to address the special characteristics of users in the humanities (Burghardt and Wolff, 2014). There are isolated approaches of integrating the whole NLP-pipeline [https://weblicht.sfs.uni-tuebingen.de/, https://hudesktop.hucompute.org/, https://textgrid.de/]. Nevertheless, the limitations are that you can recombine the provided processes in the tools but not fundamentally extend them.

We exemplify the benefits that result from the extensibility of the iLCM as follows:

Adaptability: Predefined analysis functions often require research-specific adaptations, both in preprocessing and in the analysis of text data. In the iLCM, a high degree of adaptability is provided by the ability to extensively parameterize each analysis step. Since internally each analysis method is implemented as a script written in the R programming language, it is possible to adapt the predefined methods directly. The edited processes can be reintegrated into the tool and are also available to the users of the graphical interface. In this way, iLCM can also be used to distribute tasks in interdisciplinary teams. A developer for the R components can adapt analyses for the humanities researchers and they use the processes via an easily accessible graphical interface.

Extensibility: If a needed function or method is not available in the iLCM, it should be possible to add these functions. In iLCM, new scripts can be created within the iLCM script editor to add new analysis functions or replace existing ones. Furthermore, it is possible to implement additional analyses based on intermediate results in an associated RStudio IDE.

Data export: If it is not possible or desired to fully implement a research design within the framework provided by the iLCM, it may still be possible to represent at least partial-processes using the tool. The result of these can then be exported and used in other software environments.

With the iLCM we reflect the requirements of humanists for interactive, visual interfaces and standard tools. In response to frequently necessary adjustments in the creative processing of research tasks, we have also implemented the requirements of agile development in research processes. (Heyer, Kahmann and Kantner, 2019). With this Feature we contribute to the fact that the tool used does not limit the researchers and that the openest possible processing of text-oriented research tasks is possible.

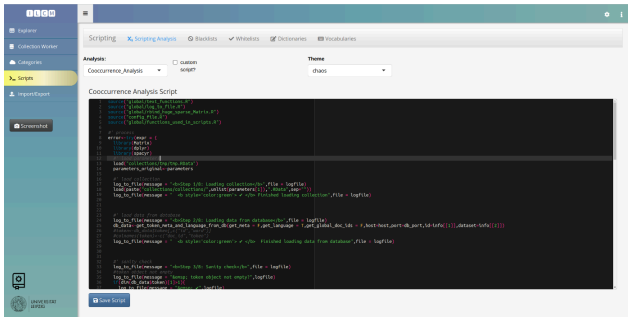


Figure 1:
The view shows the interface of the iLCM in which an editing interface for the scripts has been integrated. The example of a co-occurrence analysis shows that the underlying standard process, which can also be used via the graphical user interface, is customizable. The edited process can then be integrated again with a custom process name and is also usable for all users of the graphical interface.

Bibliography

Burghardt, Manuel, & Wolff, Christian. (2014). Humanist-Computer Interaction: Herausforderungen für die Digital Humanities aus Perspektive der Medieninformatik. Universität Regensburg. doi: 10.5283/EPUB.35716

Heyer, G., Kahmann, C. and Kantner, C. (2019) 'Generic tools and individual research needs in the Digital Humanities - Can agile development help?', *INFORMATIK 2019: 50 Jahre Gesellschaft für Informatik - Informatik für Gesellschaft (Workshop-Beiträge)*. Gesellschaft für Informatik e.V., pp. 175–180. doi: 10.18420/inf2019_ws19.

Niekler, A., Bleier, A., Kahmann, C., Posch, L., Wiedemann, G., Erdogan, K., Heyer, G., & Strohmaier, M. (2018). ILCM – A Virtual Research Infrastructure for Large-Scale Qualitative Data. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Application for visualizing and analyzing the historical network with context-centric model

Ogawa, Jun
htjk6513khbk@yahoo.co.jp
University of Tokyo, Japan

Nakamura, Satoru
nakamura@hi.u-tokyo.ac.jp
University of Tokyo, Japan

Nagasaki, Kiyonori
nagasaki@dhii.jp
International Institute for Digital Humanities

Ohmukai, Ikki
i2k@l.u-tokyo.ac.jp
University of Tokyo, Japan

Introduction

We develop the application useful for visualizing and analyzing the historical RDF data constructed on the context-centric data model, which, by introducing the new concept of *PersonInContext* (Ogawa et al., 2020), makes it possible to represent historical actors not as a unique entity, but as a collection of contextual entities.

Data Model

The basic concept is that the persons (or any other historical actors like places or organizations) should not be thought of as a single entity, being identical all the time, but as "a collection of contextual entities (Akoka et al., 2021)". Thus, we characterize our model as context-centric.

To bring this basic concept into practical use, we proposed *PersonInContext* as a new class in our ontology representing the person in a specific context. This specific context would not be defined by date information, which is not always given by historical sources, but as an interval of two historical events (Ide & Woolner, 2007). For example, in Caesar's *De Bello Gallico*, we cannot know when exactly Caesar arrived at Gaul and then defeated the Helvetians in 58 B.C.E. Still, since the former preceded the latter, we can describe Caesar in a context: from his arrival in Gaul until the defeat of Helvetians.

The advantage of this model is that it would enable us to describe historical events or relationships in a different way from previous models, such as Bio CRM (Tuominen et al., 2018) or Factoid model (Pasin & Bradley, 2013), providing that a person participating in an event is first connected to an instance of *PersonInContext*, which represents the temporal context not necessarily limited to a single event.

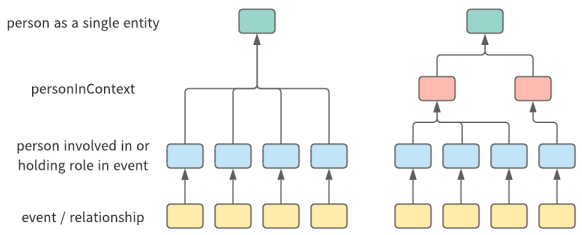


Fig.1.
Overview of our model (right)

In our model, each actor is no longer directly connected to a non-contextual entity (green square), but to a contextual one (red square). We are now able to ask a question as follows: in the context where Person_Y has a relationship with Person_X, what other relationships does Person_Y have?

Application

Based on this model, we converted the first volume of Caesar's *De Bello Gallico* to RDF data and developed an application for visualization and analysis (Data available at: https://junjun7613.github.io/RomanFactoid_v2/data.ttl). The figure below shows Caesar's ego-centric network in three different contexts. The blue node located in the center of the rightmost network is a *PersonInContext* entity representing Caesar from his arrival in the territory of Segusiavi until the end of the war against Helvetians.

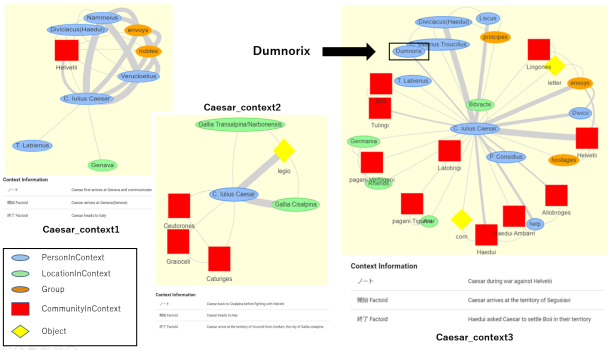


Fig. 2.
Caesar's network in different contexts

All the persons appearing in this network have their own contexts, even if this network itself represents the relationships Caesar has in one specific context. For example, if you click one blue circle node representing Dumnorix, the new window as follows pops up.

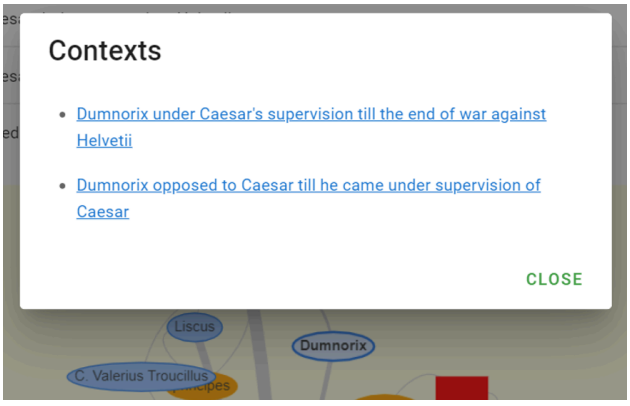


Fig. 3.
Context information of Dumnorix

Two options shown are made of *PersonInContext* entities representing Dumnorix (although these messages are now described in the data itself as character strings, it is also possible to be generated automatically). This shows that, though Caesar has several contacts with Dumnorix in his one coherent context, in terms of Dumnorix's context, he is not always the same, but is in two different contexts. Then, choosing the second option, we will move to the other network centered by Dumnorix in a certain context.

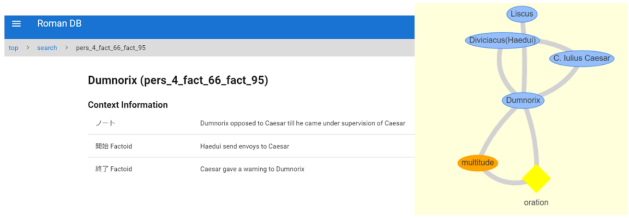


Fig. 4.
Dumnorix's network in a context

Considering this network and the previous one together, we can see the fact that, at the time Caesar had some contacts with Dumnorix during the war against Helvetians, Dumnorix had relationships with the multitude and Diviciacus in his own context. Contextual entities for Dumnorix by the way can be more than two as the definition of contexts may differ depending on the historical interpretations.

Conclusion

The application enables to visualize and analyze the change of relationships based not on the date information, which has already been applied in previous network analysis methods (Bissières, 2021), but on the contexts of historical actors mentioned in sources. Since the connectivity among historical actors is generally not clear-cut with precise dates but is highly dependent on the successive and sometimes overlapping contexts, it is necessary to deal with such contextual information in an effective way.

In this perspective, our context-centric model must introduce a useful way of representing historical information.

Bibliography

Akoka, J. et al. (2021). Conceptual Modeling of Prosopographic Databases Integrating Quality Dimensions. *Journal of Data Mining & Digital Humanities*. pp. 1-14.

Bissières, L. (2021). "Taking Time Seriously": An Empirical Approach to an American Merchant Network at the End of the 18th Century. Abstract for HNR+ResHist Conference 2021, June-July 2021.

Ide, N. and Woolner, D. (2007). Historical Ontologies. In Ahmad, Khurshid et al. (eds.), *Words and Intelligence II: Essays in Honor of Yorick Wilks*. Dordrecht. pp. 137-152.

Ogawa, J. et al. (2020). Creating a New Semantic Model for Ancient Greco-Roman Prosopography: Toward a Contextual & Historical Description of the Prosopographical Data. Abstracts for DH2020, Ottawa, July 2020.

Pasin, M. and Bradley, J. (2013). Factoid-based Prosopography and Computer Ontologies: towards an integrated approach. Literary and Linguistic Computing. Oxford: Oxford University Press.

Tuominen, J. et al. (2018). Bio CRM: A Data Model for Representing Biographical Data for Prosopographical Research. *Proceedings of the Second Conference on Biographical Data in a Digital World 2017*. pp. 59-66.

Considerations for the TEI encoding of Sino-Japanese glossed materials

Okada, Kazuhiro
k-okada@hgu.jp
Hokkai-Gakuen University, Japan

This presentation addresses considerations for TEI-based encoding of Classical Chinese texts that are glossed in Japanese from the 9th to 12th centuries (*kunten* materials; called Sino-Japanese glossed materials here). This tradition extends to nowadays, and some modern Japanese glosses have been digitally encoded (See SAT Daizōkyō Text Database Committee, 2020, for a limited encoding). Taking examples from a well-studied 9th-century glossed text, this presentation identifies barriers to encoding using an established coding framework, the Text Encoding Initiative (TEI), and discusses possible solutions. Despite their importance to Japanese philology, these materials have recently not received due attention. Identifying difficulties in encoding and establishing an encoding model for these unfamiliar materials will remove an important barrier to further study.

These materials were produced by the process of vernacular reading. Beginning in the 8th century, a body of Classical Chinese texts was read in the Japanese language. The challenges of reading these Classical Chinese texts in Japanese lie in the typological differences between the source and target languages, which glosses seek to resolve. One is a difference in the basic word order; while Classical Chinese is a subject–verb–object (SVO) language, Japanese

is an SOV language. Another is in morphology, as Japanese expresses syntactic relations with case particles and other periphrases, but Chinese largely depends on word order.

Japanese glosses are devices to record such vernacular readings on the surface of the original texts. Whereas phonogram is one of the major devices in glossing, in order not to scatter the text surface, schemata of marks borrowed from Chinese and Korean practices ("poyin" and tone glosses) were also employed. As each Chinese character generally represents a single word in the Chinese language, glosses are added to each character. The schemata set out rules for reading marks in a particular position over the surface of a character. However, the order for reading the marks is never given, even though as many as five marks may be added. Moreover, even successfully deciphering the glosses does not always produce a complete sentence, and codebreakers are forced to fill the gap. It should also be mentioned that a single manuscript may record several reading attempts that span centuries.

When encoding the glosses, it is imperative to group and relate every gloss for the same character to the others, as each gloss constitutes an intended reading. The current TEI framework offers encoding methods for most cases but has some problems due to this requirement. Samples of TEI encoding will be shown to demonstrate effectiveness of the encoding models, taking example from a 9th-century glossed text, known as the Saidaiji scroll of the *Suvarṇa-prabhāsa sutra* (or *Sutra of the golden light*), which is copied in 762 and glossed in the year around 830 (in white pigment) as well as the year 1097 (in vermilion).

The most problematic case is the treatment of marks. Currently, TEI does not have an element for marks representing a text rather than a textual signal (<metamark>). In an earlier study of the encoding of glossed materials in a specially crafted XML schema (Takada, 2010), a decoded reading of the mark was encoded. This is a possible solution because interpreting the position of a mark is to interpret what the mark represents. However, it is desirable to introduce a dedicated element for marks that the encoder wishes to leave uninterpreted, whether through customization or the expansion of the TEI framework itself. Another attempt is Yanagihara et al. (2022), where the 830 glosses of the scroll in question is decoded and encoded, with the intention of incorporating it into a diachronic corpus of the Japanese language. This nature let them encode this material as if it were written in Japanese: i.e., reordering the original text into the Japanese word order. Too, there should also be a way to preserve the original context.

As discussed earlier, preserving the original glosses means encoding the contested texts. We discuss the possible encoding methods, such as 1) <metamark>, 2) expanding abbreviation, and 3) gaiji modules. Other incompatibilities between the TEI framework and the needs of encoding glossed materials are also discussed.

Bibliography

SAT Daizōkyō Text Database Committee. (2020). Shoomangyoo gisho. https://21dzk.l.u-tokyo.ac.jp/SAT/sat_tei.html (accessed 21 April 2022).

Takada, T. (2010). Description for constructing the transcribed text of glossed material using XML. *IPSJ SIG Technical Report,* **2010-CH-85**(5): 1–8.

Yanagihara, E., Kondo, A., Takada, T. and Tsukimoto, M. (2022). Kunten shiryoo kundoku bun koopasu sakusei no igi, shuhoo, soshite kadai. Tsuuji koopasu shinpojyuumu 2022, Tachikawa: NINJAL, March 2022.

Handwritten Text Recognition and Palimpsest Analysis for Medieval Greek Manuscripts

Okada, Takashi
takashi_3.okada@toppan.co.jp
Toppan Inc.

Miyagawa, So
miyagawa.so.36u@kyoto-u.jp
Kyoto University

Kawazu, Kosei
kosei.kawazu@toppan.co.jp
Toppan Inc.

Ishii, Tatsuya
tatsuya.ishii@toppan.co.jp
Toppan Inc.

Oka, Toshio
toshio.oka@toppan.co.jp
Toppan Inc.

Fujimaki, Satoshi
satoshi.fujimaki@toppan.co.jp
Toppan Inc.

Osawa, Tomejiro
tomejiro.osawa@toppan.co.jp
Toppan Inc.

Shiki, Yoko
yoko.shiki@printing-museum.org
Printing Museum, Tokyo

Maehara, Noriko
maehara@printing-museum.org
Printing Museum, Tokyo

Ishibashi, Keiichi
keiichi.ishibashi@printing-museum.org
Printing Museum, Tokyo

Sunada, Kyosuke
ks.tkb3594@gmail.com
University of Tokyo

Nakanishi, Yasuhito
nakanishi@printing-museum.org
Printing Museum, Tokyo

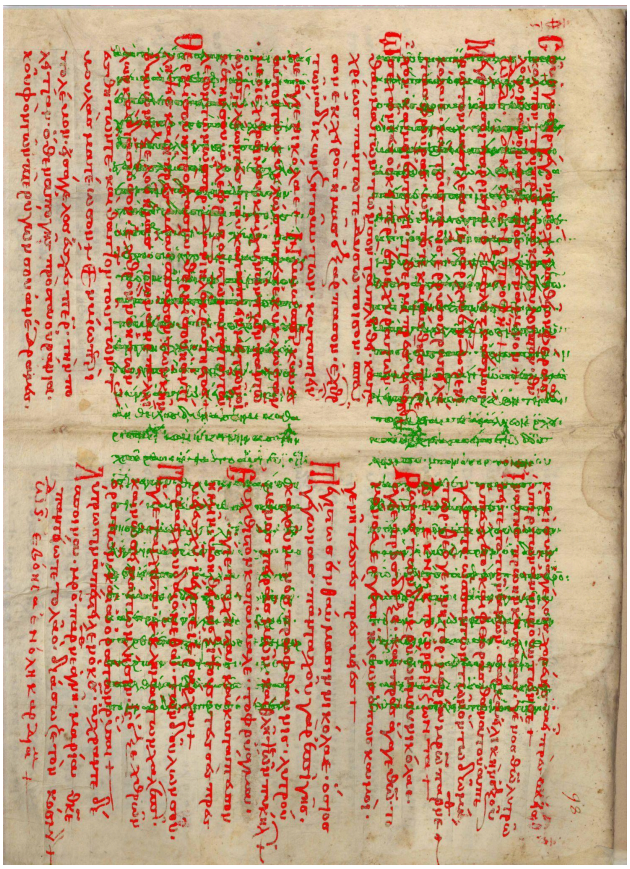
Medieval Greek Manuscript Digital Database with High-Resolution Images

Medieval Greek manuscripts were written during the Byzantine Empire in the Middle Ages. Among various Greek writing styles, medieval Greek manuscripts are known to be difficult to read due to the frequent and diverse use of ligatures and various diacritics. The Biblioteca Apostolica Vaticana (BAV) preserves many medieval Greek manuscripts that have yet to be transcribed.

Since its establishment in 2000, the Printing Museum in Tokyo (PMT) has collaborated closely with the BAV. The Cicero Project, a joint research project that commenced in 2005 and will end in 2021, aims to create an open digital database of hundreds of the BAV's medieval Greek manuscripts with high-resolution images of each folio, including palimpsests. After the project's launch, Toppan Inc., PMT's parent company, developed scanners and analysis systems.

Palimpsest Analysis

The scanner team studied page acquisition, data storage methods, and alternative methods to digitize the target pages. Then, following the decoding team's research policy, each designated palimpsest was digitized. The palimpsest analysis software superimposed the white light image and the ultraviolet image of the palimpsest digitized by the scanner. The visible text was subsequently extracted, making it easier to decipher.



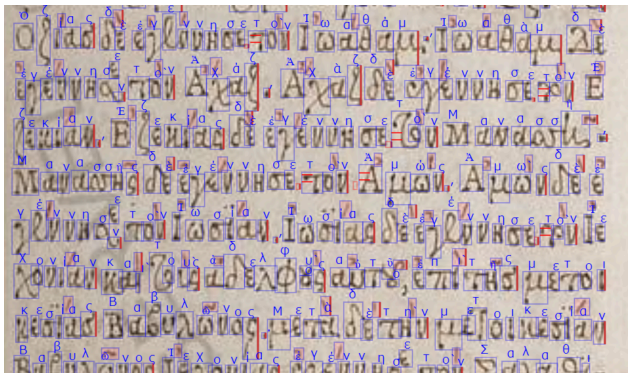
Result of the palimpsest analysis in the Cicero Project (Vat.Gr.1837)

As a result of the palimpsest analysis, a fragment of a work by 10th century Byzantine historian Leo the Deacon was found in the palimpsest manuscript Vat.Gr.1307 in the BAV collection. The fragment contained a description of Byzantine history and the origin of the Slavic peoples, which has significant differences with another already known manuscript of the same work (Janz, 2006).

Handwritten Text Recognition

It is physically impossible for the small number of researchers at the BAV to decipher approximately 30,000 images of Greek manuscripts that have been digitized. Therefore, in 2017, the new project succeeding the Cicero Project started developing a new deep-learning HTR (handwritten text recognition) system based on the *Fumi no Ha* OCR technology (Toppan Inc, 2021), originally developed for recognizing handwritten cursive early modern Japanese texts. *Fumi no Ha* includes our cursive script data set, AI cursive script recognition program, and the viewer that Toppan Printing Co., Ltd., currently known as Toppan Inc., already developed commercially. This system offers the following advantages:

- Unlike existing line-based HTR using CRNN (Shi et al., 2015; e.g., bidirectional LSTM), *Fumi no Ha* enables the identification of character coordinates for each character. As such, even when there are difficult-to-read characters, it is possible to reprint each character while referring to the image;
- Recognition results can be generated together with confidence information. Users can discard characters with low confidence levels and perform more accurate single-character recognition for them instead;
- Toppan Inc.'s web-browser-based image viewing software makes it easy to compare the original manuscript and the transcription. No special application is required for browsing; and
- The HTML format output can be displayed anywhere as long as a web browser is available. There is no requirement for dedicated systems or maintenance costs.



Training data in HTR for Medieval Greek manuscripts

Training data were prepared by character on images of handwritten manuscripts provided by the BAV (Figure 2). In this phase, two experts on medieval Greek philology directed the training data: So Miyagawa and Kyosuke Sunada. Miyagawa's experience of training Coptic OCR (Miyagawa et al., 2019) contributed to this training phase.

Fumi no Ha uses the hybrid system of character-based and line-based recognition systems. This hybrid system enables the recognition of highly complicated character layouts found in Japanese cursive manuscripts by employing the character-based system. At the same time, by using the line-based strategy, it ensures ease in the preparation of ground truth data. Medieval Greek manuscripts have various ligatures and an eccentric layout of letters. For this kind of complex character layout, the hybrid system is more appropriate than only line-based systems, such as Transkribus (Kahle et al., 2017) and OCR4all (Reul et al., 2019).

Conclusions

By taking full advantage of this set of the manuscript scanning, palimpsest analysis, and HTR programs, it is possible to build a useful database for the scholarly community, which provides high-resolution images and texts of medieval Greek handwritten manuscripts.

Bibliography

Janz, T. (2006). *Un nouveau témoin fragmentaire de Léon le Diacre le Vat*. Sofia: Centre Dujčev.

Kahle, P., Colutto, S., Hackl, G., and Mühlberger, G. (2017). Transkribus: A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. *Proc. 14th IAPR International Conference on Document Analysis and Recognition*, pp. 19–24. DOI: 10.1109/ICDAR.2017.307.

Reul, C., Christ, D., Hartelt, A., Balbach, N., Wehner, M., Springmann, U., Wick, C., Grundig, C., Büttner, A., and Puppe, F. (2019). OCR4all: An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings. *Applied Sciences* **9**(22). DOI: 10.3390/app9224853.

Shi, B., Bai, X., and Yao, C. (2015). An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. DOI: 10.1109/TPAMI.2016.264637.

Miyagawa, S., Bulert, B., Büchler, M., and Behlmer, H. (2019). Optical Character Recognition of Typeset Coptic Text with Neural Networks. *Digital Scholarship in the Humanities,* **34**(Suppl. 1): i135–i141. DOI: 10.1093/llc/fqz023.

Toppan Inc. (2021). Fumi no Ha. https://www.toppan.co.jp/biz/fuminoha/ (accessed 21 April 2022).

Uncovering the Black Fantastic: Piloting Text Similarity Methods for Finding "Lost" Genre Fiction in HathiTrust (Poster)

Parulian, Nikolaus Nova
nnp2@illinois.edu
HathiTrust Research Center, iSchool, University of Illinois, Champaign, IL USA

Dubnicek, Ryan
rdubnic2@illinois.edu
HathiTrust Research Center, iSchool, University of Illinois, Champaign, IL USA

Layne-Worthey, Glen
gworthey@illinois.edu
HathiTrust Research Center, iSchool, University of Illinois, Champaign, IL USA

Williams, Seretha
seretha.williams@augusta.edu
English and World Languages, Augusta University, Augusta, GA USA

West-White, Clarissa
whitec@cookman.edu
Carl S. Swisher Memorial Library, Bethune-Cookman University, Daytona Beach, FL USA

Magni, Isabella
isamagni@indiana.edu
HathiTrust Research Center, Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN USA

Downie, J. Stephen
jdownie@illinois.edu
HathiTrust Research Center, iSchool, University of Illinois, Champaign, IL USA

Background

Within the tradition of speculative fiction, many sub-genres of literature engage with world-building and speculation about the future of humanity, civilization, and our institutions. While the term "Afrofuturism" is commonly used to describe the ways in which artists and creators of the African Diaspora engage with the intersections of race and technology in their works, the less contested term "Black Fantastic" more accurately reflects transcultural iterations of world-building (Iton, 2008). While there is an active scholarly community studying the Black Fantastic (Third Stone, 2021), the use of computational methods in this study is inhibited by a lack of digital collections of Black Fantastic literature. Though the rise of digital libraries has led to new studies and insights into forms and histories of literary genres (Schöch, 2017; Underwood, 2016; Gittel, 2021; Wilkens, 2016), before genre-specific texts can be studied algorithmically, they must first be identified amidst the massive holdings of digital libraries, such as those of the HathiTrust. This project outlines efforts to use text similarity methods to uncover Black Fantastic texts that may be hidden thanks to incomplete metadata or to cataloging practices not conducive to fine-grained genre identification.

Methods & Analysis

Catalog searches in the HathiTrust Digital Library (HTDL) of previously-identified Black Fantastic titles revealed a preliminary list of 13 unique titles. Because metadata records for these volumes do not include "black fantastic" or "afrofuturism," common library catalog tools would not uncover these known items nor other potential Black Fantastic works. This challenge presented an opportunity to deploy technical methods, specifically text-similarity clustering, to reveal volumes of interest potentially hidden within the HTDL. We started by collecting samples of Black-authored fiction and manually labeling those identified as belonging to the Black Fantastic genre, then analyzed the lexical features that best distinguish these from both general fiction and Black-authored non-Fantastic fiction.

We randomly sampled 100 volumes each of Black-authored and general fiction, then used HTRC Extracted Features data to aggregate word usage in each volume. Our first attempt at distinguishing Black Fantastic texts necessarily relied on a very limited, 13-volume sample set. We aggregated frequently-used words most closely associated with each genre and generated a Latent Dirichlet Allocation (LDA) topic model (Blei, et al., 2003; Rehurek and Sojka, 2011) to illustrate the different sets of

characteristic words for each sub-genre. Figure 1 shows these words for each text class.



(a) General fiction (b) Black-authored fiction



(c) Black Fantastic fiction

Figure 1.

Topic model representation of distinctive words in each sub-genre of the sample. We categorized each genre with a 4-topic model. Words that most distinguish topics from the general fiction category include "duke", "car", "dollars", "president", and others, whereas topics from Black-authored fiction rely most on words such as "preacher", "church", religious", "worship", "encampments". Moreover, with the addition of words distinct to Black-authored fiction, the Black Fantastic topical signal relies more on words such as "politer", "bodkin", "sundered", "libels", "amble", "oracular", etc.

Next, we use distinct word sets to develop a classification model for the three sub-genres. We experimented with a hierarchical classification system, first attempting to distinguish Black-authored from general fiction, and then to identify the Black Fantastic sub-genre from among all Black-authored fiction. We used the Stochastic Gradient Descent (SGD) algorithm (Bottou, 2012; Pedregosa, et al., 2011) for our classification models based on Term Frequency-Inverse Document Frequency (TF-IDF) features. Figure 2 shows the confusion matrix for each classification of the test set. As we can see, the classification of Black-authored fiction works well, with 90% accuracy on average, but the Black Fantastic classification is less effective, with 67% accuracy and 75% recall. This further implies that our choice of distinctive tokens as a feature to differentiate fictional sub-genres indeed works, but to a lesser degree for detection of the Black Fantastic as a genre. These errors might result from our limited Black Fantastic training set. Beginning with more Black Fantastic texts, and identifying their most distinctive words, might improve the classification model.

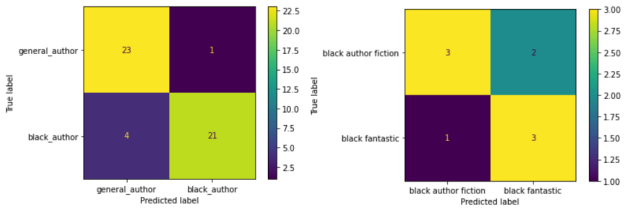


Figure 2.

Confusion matrices for a hierarchical classification model. The left-hand matrix represents classification into two collections: general and Black-authored fiction; and the right-hand represents the classification of the Black-authored set into Black Fantastic and general Black-authored works.

Future Work

Our next phase will rely on an enriched Black Fantastic dataset in order to improve our classification accuracy. We will also experiment with different features, such as full text (rather than our bag-of-words approach). In spite of limited results for our ultimate goal of discovering new Black Fantastic texts, our success with this model for distinguishing related categories gives reason for optimism given more robust data.

Note: This submission is a companion piece to another one focused on the pedagogical aspects of the "Black Fantastic" project.

Bibliography

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation. The Journal of Machine Learning Research, 3(null): 993–1022.

Bottou, L. (2012). Stochastic Gradient Descent Tricks. In Montavon, G., Orr, G. B. and Müller, K.-R. (eds), Neural Networks: Tricks of the Trade: Second Edition. (Lecture Notes in Computer Science). Berlin, Heidelberg: Springer, pp. 421–36 doi: 10.1007/978-3-642-35289-8_25. https://doi.org/10.1007/978-3-642-35289-8_25 (accessed 1 June 2022).

Gittel, B. (2021). An Institutional Perspective on Genres: Generic Subtitles in German Literature from 1500-2020. Journal of Cultural Analytics, 6(1). Department

of Languages, Literatures, and Cultures: 22086 doi: 10.22148/001c.22086.

Iton, R. (2008). In Search of the Black Fantastic: Politics and Popular Culture in the Post-Civil Rights Era. (Transgressing Boundaries: Studies in Black Politics and Black Communities). New York: Oxford University Press doi: 10.1093/acprof:oso/9780195178463.001.0001. https://oxford.universitypressscholarship.com/10.1093/acprof:oso/9780195178463.001.0001/acprof-9780195178463 (accessed 1 June 2022).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011). Scikit-learn: Machine Learning in Python. The Journal of Machine Learning Research, 12(null): 2825–30.

Rehurek, R., and Sojka, P. (2011) "Gensim–python framework for vector space modelling." NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic 3, no. 2.

Schöch, C. (2017). Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. Digital Humanities Quarterly, 011(2).

Underwood, T. (2016). Genre Theory and Historicism. Journal of Cultural Analytics, 2(2). Department of Languages, Literatures, and Cultures: 11063 doi: 10.22148/16.008.

Wilkens, M. (2016). Genre, Computation, and the Varieties of Twentieth-Century U.S. Fiction. Journal of Cultural Analytics, 2(2). Department of Languages, Literatures, and Cultures: 11065 doi: 10.22148/16.009.

About This Journal | Third Stone | Rochester Institute of Technology https://scholarwor ks.rit.edu/thirdstone/about.html (accessed 10 December 2021).

CREMMALab project

Pinche, Ariane

ariane.pinche@chartes.psl.eu
Ecole nationale des chartes, France

Within the infrastructure of the CREMMA project (Consortium for Handwriting Recognition of Ancient Materials) supported by the DIM (research funded by the Île-de-France Region) MAP (Ancient and Heritage Materials), the CREMMALab [1] project combines research questions, creation and release of data from French medieval literary manuscripts for HTR (see the *cremma-medieval* repository on Github : https://github.com/HTR-United/cremma-medieval).

The challenge of HTR is still to be taken up because of the great variety of writings, the poor readability, even for a human reader, of handwritten documents, whether due to the degradation of the source or the lack of homogeneity of the writing. Finally, the production of training data is extremely costly. However, the technical progress of the last few years in AI and neural networks allows us to produce data that significantly reduces manual work. In the Humanities, if some tools have emerged such as Transkribus (Kahle and al.), 2017) or Kraken (Kiessling, 2019) and its interface eScriptorium (Kiessling et al., 2019), developed at the EPHE, PSL university, we lack data on medieval documents to train performant models. So, producing training data is today a major issue.

The objective of the CREMMALab project is to propose open training data and HTR models for medieval documents. All data and models produced by the project are already available in the *cremma-medieval* repository on HTR-united (Chagué et al., 2021). The CREMMALab project implements transcription protocols to optimize the training of HTR models and to eventually produce homogeneous and shareable data. As a first step towards the FAIR principles, through the *cremma-medieva*l repository, we have set up some tools to ensure the citability, the durability and the quality of the data. Thus, data are described (language, date, type of document, transcription method, number of transcribed lines etc.). Then thanks to continuous integration tools, the compatibility of the XML data is checked (HTRUX), as well as the uniformity of the character sets used in the corpus (chocoMufin). Through the gathering of a corpus of medieval manuscripts, the learning process of the HTR algorithms is examined to evaluate the problems related to the constitution of training data: how to transcribe, handle abbreviations, segment words and so on. We also seek to determine thresholds of ground truth lines needed to meet quality goals, and also to evaluate the impact of the training corpus on the quality and genericity of the models.

Bicerin, an HTR model for medieval French manuscripts, is already available. It has been trained on eleven manuscripts written between the 13 th and 14 th centuries in Gothic script (about 18400 transcribed lines, see table 1). Its accuracy (CER) on this corpus is 95.38 % (dev score). This generic model, which still needs to be improved produces predictions for similar out-of-domain sources (Chantilly, Bibliothèque du Château, ms. 734, 14 e siècle) with an accuracy of 83 % (test score, see figure 1). A quick customization, requiring the addition of only two folios (336 transcribed lines) achieves an accuracy of 91 % (test score with the same sample as the first experiment, see figure 2). Most of the recognition problems come from specific difficulties due to the manuscript : word segmentation, distinction between *u* and *n*, abbreviations and reproduction quality.

The model can also be customized on a document which, at first sight, might seem incompatible with the

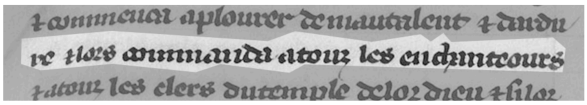
training corpus. The manuscript codex 909 from the University of Pennsylvania was written in France at the end of the 15 th century, or at the very beginning of the 16th century. It is written in *B âtarde* (this script is a hybrid of the formal style with a cursive script). Applied directly, the model shows an accuracy of 80 % (test score, figure image 3). However, *Bicerin* is very flexible, and with a customization from the addition of two folios (320 manually transcribed lines), we get an accuracy of 97 % (test score with the same sample as the precedent experience, see figure 4). The high recognition of this writing is certainly related to the regularity of the writing and the quality of the support and the reproduction, two criteria that seem to be more important for HTR than the type of writing.

In the future, we hope to increase the number and diversity of our training data to improve the genericity of the model and its robustness. We also hope to determine its breakpoints and to delimit contexts in which a generic model is enough and those in which it is relevant to create a personalization or another model from scratch.

Annexes

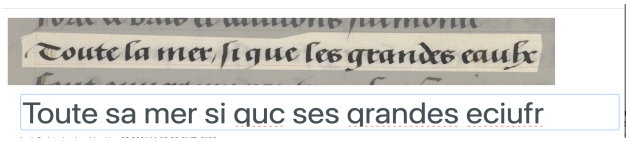
Table 1 : Cremma medieval corpus

Manuscript	Date	Transcribed Lines
BnF, ms fr. 17229	13th	161
BnF, ms fr. 13496	13th	159
BnF, ms fr. 411	14th	153
BnF, Arsenal 3516	13th	1991
BnF, ms fr. 22549	14th	411
BnF, ms fr. 24428	13th	1295
BnF, ms fr. 412	13th	4551
BnF, ms fr. 844	13th	1026
Cologny, bodmer, 168	13th	1927
Vaticane, Reg. Lat., 1616	14th	1726



re nlors counmnanda a touz les eu chiceours

Bicerin model prediction on Chantilly, ms. 734



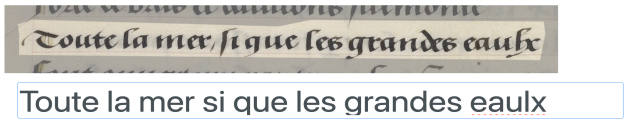
Toute sa mer si quc ses qrandes eciufr

Bicerin model prediction on Philadelphy, university of pennsylvania, ms codex 909



re ⌐lors coummanda atouz les en chanteours

Finetuned model prediction on Chantilly, ms. 734



Toute la mer si que les grandes eaulx

Finetuned model prediction on Philadelphy, university of pennsylvania, ms codex 909

Bibliography

Bulacu, M., & Schomaker, L. (2007), "Automatic Handwriting Identification on Medieval Documents", *14th International Conference on Image Analysis and Processing (ICIAP)*, 2007, 279-284, <https://doi.org/10.1109/ICIAP.2007.4362792>.

Chagué, A., Clérice, T., & Chiffoleau, F. (2021), *HTR-United, a centralization effort of HTR and OCR ground-truth repositories mainly for French languages*, <https://github.com/HTR-United/htr-united> (Original work published 2020)

Dome, S., & Sathe, A. P. (2021), "Optical Charater Recognition using Tesseract and Classification", *International Conference on Emerging Smart Computing and Informatics (ESCI)*, 2021, 153–158. https://doi.org/10.1109/ESCI50559.2021.9397008

Fischer, A., Wuthrich, M., Liwicki, M., Frinken, V., Bunke, H., Viehhauser, G., & Stolz, M. (2009). "Automatic Transcription of Handwritten Medieval Documents", *15 th International Conference on Virtual Systems and Multimedia*, 2009, 137–142, <https://doi.org/10.1109/VSMM.2009.26>.

Kahle, P., Colutto, S., Hackl, G., & Mühlberger, G. (2017). "Transkribus—A Service Platform for Transcription, Recognition and Retrieval of Historical Documents", *14 th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, 4, 19–24, < https://doi.org/10.1109/ICDAR.2017.307>.

Kestemont, M., Christlein, V., & Stutzmann, D. (2017), « Artificial Paleography: Computational Approaches to Identifying Script Types in Medieval Manuscripts. *Speculum »*, 92(S1), S86–S109. https://doi.org/10.1086/694112

Kiessling, B. (2019), "Kraken—An Universal Text Recognizer for the Humanities", *DH2019*, Utrecht, <https://dev.clariah.nl/files/dh2019/boa/0673.html>.

Kiessling, B., Tissot, R., Stokes, P., & S toekl, D. (2019), "eScriptorium: An Open Source Platform for Historical Document Analysis", *International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2019, 19–19, <https://doi.org/10.1109/ICDARW.2019.10032>.

Pinche, A., & Clérice, T. (2021), "HTR-United/cremma-medieval: 1.0.1 Bicerin (DOI)", *Zenodo,* < https://doi.org/10.5281/zenodo.5235186>.

Ströbel, P. B., Clematide, S., & Volk, M. (2020), « How Much Data Do You Need? About the Creation of a Ground Truth for Black Letter and the Effectiveness of Neural OCR », *Proceedings of the 12th Language Resources and Evaluation Conference*, 3551–3559. https://aclanthology.org/2020.lrec-1.436

Notes

1. Presentation project : < https://cremmalab.hypotheses.org>

Buddhist Murals of Kucha on the Northern Silk Road. An Approach to Semi-Automated Annotation

Radisch, Erik
radisch@saw-leipzig.de
Saxon Academy of Sciences and Humanities in Leipzig, Germany

The Buddhist cave complexes in the region of Kucha, located on the northern Silk Road (in the Xinjiang Uyghur Autonomous Region, People's Republic of China) house impressive wall paintings dating approximately from the 5th to 10th centuries. The first evidence of a past Buddhist culture was discovered at the beginning of the 20th century, after which several countries sent expeditions to the area to investigate the religion that had once been dominant in the region. It was a sensation when various Buddhist cave complexes were discovered, the largest of which included as many as 400 caves. At that time, the first photographs of the actual state of the caves were taken and pieces of the paintings were extracted from the caves and transferred to the respective national museums. Sales and losses due to war led to the fact that nowadays fragments of the murals are spread all over the world, making it very difficult to assign it to the individual caves of origin (Further information: Yaldiz 1987; Popova 2008; Dreyer 2015).

The project presented here has taken on the task of documenting and describing the murals in situ and the individual pieces available worldwide and, with the aid of historical photographs, of virtually reinserting them into their original context.[1]

The project makes use of modern possibilities of the Digital Humanities in that not only an extensive textual description of individual scenes is carried out, but also the pictorial contents of the repetitive representations are recorded and enriched with digital methods. For this purpose,the digital image annotation tool Annotorious[2] is used to directly annotate the content with a taxonomy comprising about 1,000 entries.

While annotating objects in the image makes it possible to ensure scientific traceability of identified objects, it is also a very extensive and time-consuming task. Many objects have to be annotated repeatedly because they appear in many images in different contexts. Also, there are sometimes several images of an object from different perspectives available or there are images from the time of the expeditions where detached parts can still be seen in their original context. So there is the necessity to annotate sometimes very similar or same objects several times. However, transferring annotations is difficult. Even if photographs of the same objects are available, different viewpoints and different lenses may cause the photographs to be distorted. It is hardly possible to perform this task automatically using conventional computer vision methods.

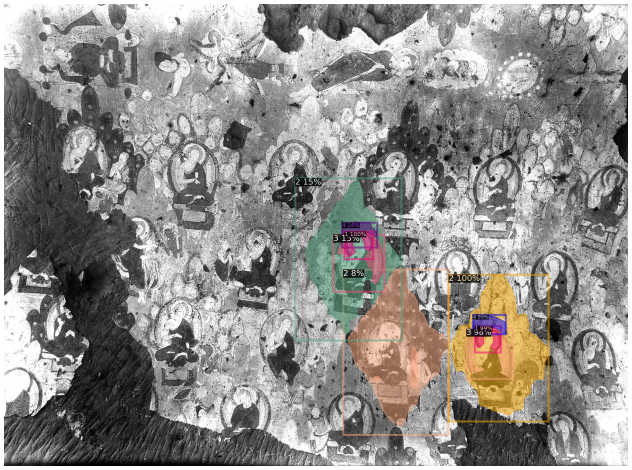
For this reason, the project is currently attempting to train region based convolutional neural networks (RCNNs)[3] using the annotations already made, in order to be able to perform at least parts of the annotation semi-automatically in the future.

So far, RCNNs have been used in the Digital Humanities mainly to identify, locate and order objects in images (see for example: Howanitz et al. 2019; Arnold/Tilton 2019; Duhaime 2019; Helm et al. 2021). Their use for semi-automated annotation has not been implemented, at least to the knowledge of the author of this poster proposal. It may also be a risky endeavour, since the edges of such detected regions often fray and not the whole object is detected. However, it may be worth a try, if only to test the limits of the application of such systems in the Digital Humanities.

The conditions in this project are very good. Nearly 8,000 polygons already exist, which have been used in a total of nearly 10,000 annotations (a polygon can be linked

to several elements of the taxonomy). Some objects have been annotated over 500 times. However, there are also some problems to be considered. For example, there are two fundamentally different types of imagery. On the one hand there are photographs (historical and modern) and on the other hand there are drawings of some paintings. Since these categories of images are very different, they were also separated for the training. It can be assumed that the recognition on photographs should be clearly more difficult, since here paintings are often in very bad condition and are to be identified only with difficulty even by an experienced eye.

A first attempt has been encouraging, because not only some objects were found but for example some of the nimbi around the head were drawn very well. (See figure 1) The poster will present the results (failure or success) of the experiments that will be performed in the next months. It should invite to inform about the project, to share experiences about the use and training of neural networks and to encourage their use.



Example photograph for automated annotation. First number: taxonomy-key, second number: confidence. By far not all possible annotations were detected Nimbi where detected good, rhombi need still further improvement.

Bibliography

Arnold, T. and Tilton, L . (2019). Distant viewing: analyzing large visual corpora, Digital Scholarship in the Humanities 36(1), DOI: 10.1093/digitalsh/fqz013.

Dreyer, C. (2015). *Abenteuer Seidenstraße: Die Berliner Turfan-Expeditionen 1902–1914*. Leipzig: Seemann.

Duhaime, D. (2019). PixPlot. https://github.com/YaleDHLab/pix-plot

Helm, W., Schmideler, S., I m, C., Mandl, T., Kollmann, S. and Müller, L. (2021). Wie sich die Bilder ähneln. Vom Zufallsfund zur systematischen Forschung im Bereich der automatisierten Bildähnlichkeitssuche. In: Burghardt, M., Dieckmann, L., Steyer, T., Trilcke, P., Walkowski, N.-O., Weis, J. and Wuttke, U. (2021). Fabrikation von Erkenntnis: Experimente in den Digital Humanities. DOI: 10.26298/melusina.8f8w-y749-wsdb.

Howanitz, G., Bermeitinger, B., Radisch, E., Gassner, S., Rehbein, M. and Handschuh, S. (2019, July 11). Deep Watching - Towards New Methods of Analyzing Visual Media in Cultural Studies. Digital Humanities 2019 (DH2019), Utrecht, Netherlands. https://dev.clariah.nl/files/dh2019/boa/0335.html.

Konczak-Nagel, I. a nd Zin, M. (2020). Essays and Studies in the Art of Kucha (Leipzig Kucha Studies 1). New Delhi: Dev Publishers.

Popova, I. F. (ed.) (2008), *Russian Expeditions to Central Asia at the Turn of the 20th Century: Collected Articles*. St Petersburg: Slavia Publishers.

Wu, Y., Kirillov, A., Massa, F., Lo, W. and Girshick, R. (2019). Detectron2, https://github.com/facebookresearch/detectron2.

Yaldiz, M. (1987). *Archäologie und Kunstgeschichte Chinesisch-Zentralasiens (Xinjiang)*. Leiden: Brill, Handbuch der Orientalistik, Abteilung 7, Kunst und Archäologie, Band 3, Innerasien.

Notes

1. https://kucha.saw-leipzig.de; https://www.saw-leipzig.de/de/projekte/wissenschaftliche-bearbeitung-der-buddhistischen-hoehlenmalereien-in-der-kucha-region-der-noerdlichen-seidenstrasse/introduction/kucha-murals
2. https://recogito.github.io/annotorious/ The Project has its own series «Leipzig Kucha Studies». The fist Book is: Konczak-Nagel/Zin 2020
3. The project uses for this purpose detectron2 https://github.com/facebookresearch/detectron2

Pragmatic Research Data Management in the Humanities: Dark and Cold Archiving at the Data Center for the Humanities

Rau, Felix
f.rau@uni-koeln.de

Data Center for the Humanities (DCH), University of Cologne, Germany

Helling, Patrick
patrick.helling@uni-koeln.de
Data Center for the Humanities (DCH), University of Cologne, Germany

Barabucci, Gioele
gioele.barabucci@ntnu.no
Department of Design Faculty of Architecture and Design, Norwegian University of Science and Technology, Norway

Introduction

Centers for Research Data Management (RDM) rightly focus on making research data findable, accessible, interoperable, and reusable in the sense of the FAIR Principles (Wilkinson et al. 2016). Repositories for storing research data in a FAIR way are of fundamental importance (Mathiak et al. 2019).

However, data protection or copyright can impose restrictions on making data FAIR. The software and workflows commonly used by RDM centers are often unable to accommodate such particular needs, or, when they do, require expensive manual interventions.

There is thus the need to find pragmatic solutions to provide safe, state-of-art archival workflows for data that follows good scientific practice (DFG 2019) but is currently unable to comply with all mandates of the FAIR principles

This poster introduces the archiving workflow established at the Data Center for the Humanities (DCH) at the University of Cologne for those cases where accessibility is not desired, and findability is optional. [1] This workflow allows us to archive research data in a structured, automated way on tape-storage from infrastructure providers at the University. The research data is thus sustainably stored but not findable and accessible (dark archiving) or, optionally, its metadata is published and findable (cold archiving).

Background

The DCH is a central institution at the Faculty of Arts and Humanities of the University of Cologne. DCH supports and advises researchers within the Humanities in RDM questions over the entire research data life cycle (Blumtritt et al. 2018, Helling et al. 2018). [2] While the DCH is responsible for domain-specific RDM in the Humanities, central institutions of the university offer basic IT services for RDM, which are used by the DCH.

Implementation of the archiving workflow

The University of Cologne does not provide an institutional archive and there are no regional, national, or supranational cold and/or dark archiving solutions available to the researchers of the Faculty of Arts and Humanities.

To meet the demand articulated by the faculty's researchers, we implemented a dark archiving service. In designing the service, we focused on sustainability, in particular in making it maintainable with a minimal amount of person hours. To achieve this, we developed an archival workflow around standards and technologies that could easily be implemented using software and hardware that is already available at the University and maintained by the IT department. [3]

All the software code that supports the workflow is written in the form of short Bash scripts. Bash was chosen for its ubiquity, stability, and for being a well understood technology (many experienced users and a wealth of published information).

The archiving service comprises four key steps: 1) preparation of archival packages, 2) ingestion via university's long-term tape archive, 3) creation of human-readable descriptive metadata, and 4) optionally, publication of metadata. Mirroring this workflow, we also implemented a retrieval workflow.

The archival packages are prepared by our software by laying out the received research data according to the BagIt conventions and adding BagIt metadata. BagIt (Kunze et al. 2018) defines a simple way of packaging digital content with data integrity checks and metadata. The bag also contains a structured readme file with basic information provided by the researcher.

The BagIt-compliant archival packages are ingested into the IBM Tivoli Storage Manager robotic tape library maintained by the IT department. [4] The archival workflow guidelines describe how to map the research data files to the structure required by Tivoli. This mapping is implemented in code and is performed automatically.

Once the data is archived, descriptive metadata with core information about the data and how to access it is generated and made available for internal use.

Regarding the requirements of the researchers, we later extended the service with a cold archiving process with published and findable metadata. When that is desired, the core information is published as a webpage via the university's Typo3 CMS. Structured dataset metadata is embedded as Schema.org/JSON-LD. Additionally, a

DataCite metadata file is created from the readme file and is used to register a digital object identifier (DOI) via the DOI service of the university library. [5]

Conclusion

Our archiving service provides state-of-the-art dark and cold archiving for research data that cannot be published as FAIR data. It achieves its goals in pragmatic ways: it saves manual labor by being fully automated, and it ensures the sustainability of the process by making use of existing, widely available and well maintained technologies.

Bibliography

Blumtritt, J., Helling, P., Mathiak, B., Rau, F. and Witt, A. (2018) Forschungsdatenmanagement in den Geisteswissenschaften an der Universität zu Köln. In: *o-bib, Das offene Bibliotheksjournal*, p. 104-17. DOI: 10.5282/o-bib/2018H3S104-117.

DFG, Deutsche Forschungsgemeinschaft (2019). Guidelines for Safeguarding Good Research Practice. Code of Conduct. DOI: 10.5281/zenodo.3923601.

Helling, P., Blumtritt, J. and Mathiak, B. (2018). Der Beratungsworkflow des Data Center for the Humanities (DCH) an der Universität zu Köln. In: *o-bib, Das offene Bibliotheksjournal*, p. 248-61. DOI: 10.5282/o-bib/2018H4S248-261.

Kunze, J., Littman, J., Madden, E., Scancella, J. and Adams, C. (2018). RFC8493 "The BagIt File Packaging Format (V1.0). *Internet Engineering Task Force*. Online: [last request: 08th of December 2021]; DOI: 10.17487/RFC8493.

Mathiak, B., Metzmacher, K., Helling, P. and Blumtritt, J. (2019). The Role Of Data Archives In The Humanities At The University Of Cologne. *DH 2019 Conference*, 8-12 July 2019, Utrecht University. DOI: 10.34894/geqeko.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Bonino da Silva Santos, L., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. In: *Scientific Data 3*, Article number: 160018. DOI: 10.1038/sdata.2016.18.

Notes

1. https://dch.phil-fak.uni-koeln.de/ [last request: 26th of November 2021].
2. https://dch.phil-fak.uni-koeln.de/fdm-services [last request: 26th of November 2021].
3. https://rrzk.uni-koeln.de/en/ [last request: 08th of December 2021].
4. https://rrzk.uni-koeln.de/en/data-storage-and-share/backup-system-tsm [last request: 08th of December 2021].
5. https://schema.datacite.org/ [last request: 08th of December 2021].

Modeling the Multivalent Perspectives of US Immigrant Narrative

Rodrigues, Elizabeth Sarah
rodrigue8@grinnell.edu
Grinnell College, United States of America

This poster details an exploratory project to model late-nineteenth and early-twentieth century US immigrant life narratives using sentiment analysis. The immigrant narrative of arrival, struggle, and achievement (of economic gain, cultural citizenship, and/or legal status) is one of the central mythologies of United States identity, and autobiographical texts are one of the central literary sources of that mythology. Literary scholars have long sought to complicate this simplistic but persistent narrative, arguing that both fictional and autobiographical immigrant narratives offer "portraits of ambivalent relationships to U.S. nationalism [that] suggest that we cannot refer to a single politics" of "celebratory, accusatory, documentarist… [or] ironized declarations of belief in the nation" but should instead attend to these texts' "interwoven registers of seemingly inconsistent perspectives on the nation" (Miller 2014). My project seeks to develop a model for encoding and visualizing late-nineteenth and early twentieth-century US immigrant autobiographical narratives' that recognizes and surfaces such "multivalent perspectives" (Miller) on the US nation. Put another way, developmental and linear conceptions of form in immigrant narrative would seek to map a subject's progress from departure to arrival to

citizenship; from economic precarity to productive labor to material stability; from national otherness to national identity. A multivalent conception would recognize and represent oscillating statuses—affective, material, legal, and others—throughout the narrative frame. Developing a multivalent approach to modeling immigrant narrative intersects with broader digital humanities efforts to model plot in literary texts. Sentiment analysis also offers a level of granularity that could be useful for approaching the multivalent perspectives of the immigrant narrative. Rather than defining plot through a handful of landmark events, analysis at the level of the word/sentence has the potential to capture shifting affective relationships to the idea and reality of the US nation.

The first stage of this project will be to examine selected fully transcribed texts using Matthew Jockers's Syuzhet, a sentiment analysis toolkit attempting to visualize plot built in R. My explorations will be guided by the validation procedures outlined by Elkins and Chun (2019) in their study a modernist novel. Models of these initial texts will be used to identify key plot points and closely examine these sections of the text for recurring terms or themes. Identification of recurring terms in pivotal affective moments will allow me to begin customizing the underlying lexicon. As Kim and Klinger (2021) note, lexicon-driven approaches to sentiment analysis have predominated literary research uses because of the high value humanistic inquiry places on transparency of algorithmic operations for the purposes of validation. Ultimately, I would hope to be able to use this custom lexicon to get beyond canonical texts and examine plots in a broader corpus of US immigrant life writing during this period. Limitations of the project include: 1) all of the texts examined will be in English (inevitably skewing the selection of texts towards authors who are writing for an English-speaking audience and perhaps more likely to self-censor or project more developmental, assimilationist narratives); 2) the corpus is for now heavily reliant on a single source, Louis Kaplan's *Bibliography of American Autobiographies*; 3) optical character recognition error rates will have to be considered for texts that do not have transcriptions, which is most of them; and 4) existing sentiment analysis lexicons may not be properly weighted to account for vocabulary most relevant to these texts. The project seeks to begin to mitigate this final limitation by beginning to develop a custom lexicon based on this exploratory work.

Bibliography

Elkins, K. and Chun, J. (2019). Can Sentiment Analysis Reveal Structure in a Plotless Novel? Accessed 10 April 2022: https://arxiv.org/abs/1910.01441.

Kim, E. and Klinger, R. (2021). A Survey on Sentiment and Emotion Analysis for

Computational Literary Studies. Accessed 10 April 2022: https://arxiv.org/abs/1808.03137.

Miller, J. L. (2014). The Immigrant Novel. In Wald, P. and Elliott, M.A. (eds), The Oxford History of the Novel in English. Oxford, Oxford University Press, pp. 200-217. Published online March 2015: 10.1093/acprof:osobl/9780195385342.003.0013.

Spatio-Temporal Analysis of the Dutch East India Company Archive through Paper Watermarks

Sakamoto, Shouji
sakamoto@mac.com
Ryukoku University, Japan

1. Introduction

Digital archives of paper-based cultural properties, such as documents, printings, provide digital images and bibliographic data, but they do not usually provide material data such as paper, ink. Various paper features, such as paper fiber and watermark, which are useful data in the study of paper-based cultural properties, are recently being extracted from paper (Sakamoto et al., 2013, 2016; Sakamoto, 2020; Johnson, 2018; Johnson et al., 2021;).

The Dutch East India Company (VOC) archives, consisting of around 20,000 items (or books), are held by the Nationaal Archief, Den Haag. In particular, the archive concerning the VOC Japan factory consists of 1,952 items (Roessingh, 1964). In this study, 220 items on VOC Japan factory from 1614 to 1831 are investigated. Paper diversity as well as change of paper are studied, depending on periods, through spatiotemporal analysis. The reasons for the change in paper are also discussed.

2. Methodology

Beta radiography is useful in the making of a watermark and a laid pattern (Ash, 1998); however, it is not easy to prepare the equipment, and the exposure times are longer than several hours (Rakvin et al., 2014; Boyle et al., 2009). Therefore, white LEDs were adopted as a light source along with a digital camera to take digital images of the paper under transmitting light.

3. Results and Discussions

Among the 220 items, over 20 watermarks, such as *Strasburg Fleur de Lis, Pro Patria, Arms of Amsterdam* were found, and there were at least 49 countermarks, such as the papermaker's initial, company name, and so forth.

Such watermarks provide information on the location of the production, paper mill, and papermaker's name (Nicolaï, 2005; Heawood, 2003; Churchill, 1965). As the VOC archive usually has date information (year, month, day), spatiotemporal analysis is possible.

Result 1: In the period between 1610 and 1663, Japanese *torinoko* paper (high-quality *gampi* paper) was used for archival purposes (Yasuda, 2016; Sakamoto et al., 2022).

According to VOC records around 1660, VOC spent too much (40,000-50,000 gulden/year) for archival paper and decided to establish the paper mill in Batavia (Landwehr, 1994) to cut the cost of paper. At that time, *torinoko* paper was more expensive than western paper, which may have been the reason for not adopting *torinoko* paper for the VOC archival.

Result 2: In the period between the 1640s and 1710s, mainly French paper, which was made in the Angoumois region, was used, except for the Japanese *torinoko* paper mentioned above.

Initials AJ and GVH stand for the Dutch factories Abraham Janssen and Gills van Hoven, respectively. Abraham Janssen managed several paper mills in Angoumois in the period between 1635 and 1710, and the paper produced was exported to neighboring countries, such as England and Holland (Churchill, 1965).

Result 3: In the period between the 1730s and 1790s, mainly Dutch paper, which was made in the Zaan region, was used.

The change from French paper to Dutch paper is attributed to the religious war in France (Edict of Fontainebleau etc.) and the invention of Hollander beater in Holland. Many Huguenots (French protestant) craftsmen, including papermakers, moved to other countries such as England and Holland [the decline of French paper] (Scoville, 1967; Rosenband, 2000). On the other hand, Holland accepted many immigrants, including the Huguenots, and grew to be an industrial country. Especially, in the Zaan region located north of Amsterdam, paper mills were established by Pieter Van der Ley, Jan Honig, and so on. Zaan paper was exported to neighboring countries [the rise of Holland paper].

Result 4: In the period between the 1800s and 1820s, types of Dutch (Zaan and Veluwe), French, American, English, Japanese *kozo* paper were uncovered, and probably Batavian paper, as well.

After the French Revolution, Holland was controlled by France, and VOC ceased to exist after 1799. At that time, no ship from Holland visited Batavia; therefore, the VOC Batavia factory hired mainly American ships to continue trading (1798-1809) in Asia (Kanai, 1966). Subsequently, Batavia was occupied by England (1811-1816). This chaotic situation broke the stable supply of paper from Holland. As a result, different types of paper from various countries exist in that period.

4. Concluding Remarks

The results obtained from paper feature analysis are completely new knowledge, which is impossible to derive from usual digital archive data. However, the results are based only on 220 items, and thus more accurate results are likely to be uncovered by the investigation of additional items. Consequently, paper analysis data are useful in the study of paper-based cultural properties in digital archives.

Bibliography

Ash, N. E. (1998). Watermarks in Rembrandt's Prints, National Gallery of Art.

Boyle, R. D. and Hiary, H. (2009). Watermark location via back-lighting and recto removal, International Journal of Document Analysis and Recognition (IJDAR), 12(1): 33–46.

Churchill, W. A. (1965). Watermarks in Paper in Holland, England, France, Etc., in the XVII and XVIII Centuries & Their Interconnection, MENNO HERTZBERGER & CO.

Heawood, E. (2003). Watermarks, Mainly Of The 17th And 18th Centuries, Martino Pub.

Johnson, C. R. (2018). WImBo — Watermark imaging box project: A digital art history data acquisition tool, 2018 52nd Annual Conference on Information Sciences and Systems (CISS), IEEE.

Johnson, C. R., Sethares W. A. and Ellis, M. H. (2021). Overlay Videos for Quick and Accurate Watermark Identification, Comparison, and Matching, Journal of Historians of Netherlandish Art, 13:2.

Kanai, M. (1966). Japan-Netherlands trade in the period of chartering neutral vessels (中立国傭船期の日蘭貿易), Journal of the Japan Society for Nautical Research, 8: 55-68. (in Japanese)

Landwehr, J. (1994). De Boekenwereld, 11: 261-263.

Nicolaï, A. (2005). Histoire des Moulins à papier du Sud-Ouest de la France (1300-1800) - tome 1, REGIONALISMES.

Rakvin, M., Dragojević, A. and Markučič, D. (2014). Application of Computed Radiography (CR) for characterization of historical documents, 11th European Conference on Non-Destructive Testing (ECNDT 2014).

Rosenband, L. N. (2000). The Competitive Cosmopolitanism of an Old Regime Craft, French Historical Studies, 23(3): 455-476.

Roessingh, M. P. H. (1964). Inventaris van de archieven van de Nederlandse Factorij in Japan te Hirado [1609-1641] en te Deshima, [1641-1860], 1609-1860, Nationaal Archief, Den Haag.

Sakamoto, S. and Okada, Y. (2013). Paper Analysis and Paper History from Ancient Chinese Paper to Japanese Washi, The International Conference on Culture and Computing 2013, 51-56.

Sakamoto, S. and Oda, H. (2016). Paper Analysis of East Asian Historical Documents with a focus on Paper Laid and Chain Lines, IPSJ SIG Technical Report, 2016-CH-111: 1-8. (in Japanese)

Sakamoto, S. (2020). Paper Feature Extraction from Digital Images, Digital Humanities 2020, http://dx.doi.org/10.17613/4p5t-tn15

Sakamoto, S. and Vilmont, L.V. (2022). Japanese and Chinese Paper in Rembrandt Etchings, The Future of Tradition in the Arts, East and West, Proceedings of the ICDAD, ICFA, and GLASS International Committees at the ICOM Kyoto 2019 General Conference. (in press)

Scoville, W. J. (1967). Government Regulation and Growth in the French Paper Industry During the Eighteenth Century, The American Economic Review, 57(2): 283-293.

Yasuda, T. (2016). Study on "Oriental Papers" in the Archives of the 17th century Dutch East India Company, Study of Washi Culture, 24: 20-51. (in Japanese)

CanAICompose? A web-based tool for deep music generation

Schlör, Daniel
schloer@informatik.uni-wuerzburg.de
University of Wuerzburg

Hotho, Andreas
hotho@informatik.uni-wuerzburg.de
University of Wuerzburg

Introduction

When it comes to creating new music, the composer as creative and well educated human being is undoubtedly the best source of interesting, well sounding and touching music, as perceiving sounds as music is very natural and subjective for humans. Consequently the definition and judgment of the aforementioned attributes is typically subjective and thus hard to formalize as required to formulate the question in a Digital Humanities driven approach. Nevertheless have people encountered the art of composing music from an algorithmic view very early as for example musical dice games which were invented by Mozart and contemporaries, which combines previously composed phrases to random patterns (Ruttkay 1997) and rules for example for harmonizing chorales in the style of J.S. Bach have been formalized (Ebcioğlu 1990). From a more general perspective however, the formalization of rules or knowledge is a difficult and tedious task and resulting compositions are limited by the knowledge or rules and surprising or genre breaking compositions are not to be expected (Papadopoulos & Wiggins 1999). Besides several other approaches to algorithmic composition, the family of Artificial Neural Networks (NN) have been well studied (Ji et al. 2020) and with the availability of high performance computing hardware, larger model architectures such as Transformers, which have shown for text-generation tasks near-human performance (Radford et al. 2020) and learn from examples, have been recently introduced to generate polyphonic music (Huang et al. 2018, Wu & Yang 2020).

These models are computationally complex and have tight hardware requirements, which hinder potential applications as supporting composition tool as well as playful interaction to explorethe cognitive creative process of composition (Gardner 1982) and the limitations of AI in this context.

We therefore introduce a web based tool available at https://go.uniwue.de/canaicompose, to make complex transformer based models more accessible to the general public by providing a simple web-interface to query the models. With the opportunity to rate individual pieces, we include a feedback mechanism to collect large-scale annotations and metadata, which we want to use to evaluate, which characteristics potentially pleasing synthetically generated music share and refine our models according to these characteristics.

Corpus

We restricted our corpus to piano pieces by classical composers which are included in the large collection of MIDI files curated by kunstderfuge.com and the maestro data set (Hawthorne et al. 2019).

Approach

Our NN model uses the Transformer architecture (Vaswani et al., 2017) including relative positional self attention as proposed by (Shaw et al. 2018) which has proven to yield promising results for polyphonic music generation with long-term structure (Huang et al. 2018). The MIDI files are encoded with note-on, note-off, time-shift and velocity events following (Oore et al. 2020).

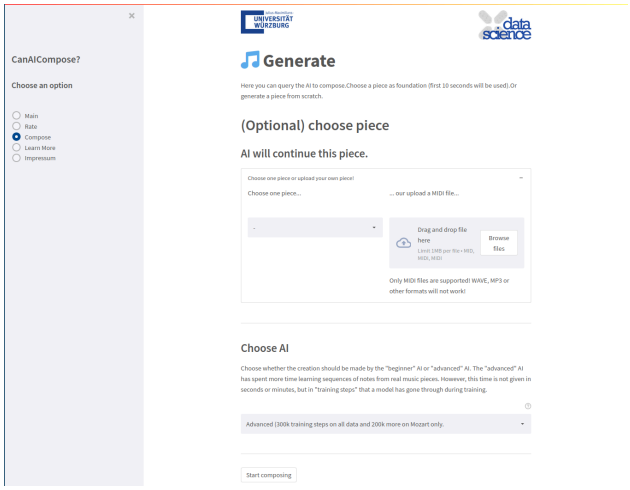


Figure 1:
Screenshot of our CanAICompose tool

The Tool

Our tool is divided in two parts, a back-end serving the machine learning models and a front-end serving a web-interface (see Figure 1) to interact with the NN models. The back-end currently allows the querying of two different models, one model trained on compositions of all composers and one model which was additionally fine-tuned on compositions by Mozart, to allow an evaluation, to which extend a well performing model can be constrained to generating music similar to a certain style which for example a composer using this tool might aspire.

The back-end has the functionality to generate a new piece from scratch or to be primed by a user-predefined snippet, which the model is meant to continue and can be queried by API commands to be adaptable for a seamless integration in sheet music notation or composition tools.

The front-end was built upon the streamlit.io library for easy integration of NN models in a web-interface and was constructed with the general public as user in mind exploring the possibilities of NN generated music and rating the pieces subjectively. It therefore refrains from displaying symbolic notation such as sheet music but allows to play, rate and download the generated material. The code for this project is available here.

As consequent work, we want to fine-tune different NN models on various epochs of classical music and incorporate a classification model to predict the most promising model for a given preconditioning snippet. Based on the subjective ratings collected with this web-app, we want to incorporate a filtering algorithm to evaluate rejected or preferred music which allows musicologists to analyze the quality of deep generated music with respect to musical factors such as tension (Lerdahl, 1996), or melodic complexity (Narmour, 1992).

Bibliography

Bharucha, J. J., and Todd, P. M. (1989). Modeling the perception of tonal structure withneural nets. *Computer Music Journal*, 13.4: 44-53.

Ebcioğlu, K. (1990). An expert system for harmonizing chorales in the style of JS Bach. *The Journal of Logic Programming*, 8.1-2: 145-185.

Gardner, H. (1982). Art, Mind, and Brain: A Cognitive Approach to Creativity. *Perseus Books Group*

Hawthorne, C. , Stasyuk, A., Roberts, A. et al. (2019). Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset. In I *nternational Conference on Learning Representations*.

Huang, C. A. et al. (2018). Music transformer. *arXiv preprint arXiv:1809.04281.*

Ji, S., Luo J., and Yang, X. (2020). A Comprehensive Survey on Deep Music Generation: Multi-level Representations, Algorithms, Evaluations, and Future Directions. *arXiv preprint arXiv:2011.06801.*

Lerdahl, F. (1996). Calculating tonal tension. Music Perception, 13.3: 319-363.

Narmour, E. (1992). The analysis and cognition of melodic complexity: The implication-realization model. *University of Chicago Press*.

Oore, S., Simon, I., Dieleman, S. et al. (2020). This time with feeling: learning expressive musical performance. *Neural Comput & Applic*, 32, 955–967

Papadopoulos, G., and Wiggins, G. (1999) AI methods for algorithmic composition: A survey, a critical view and future prospects. *AISB symposium on musical creativity,* Vol. 124.

Radford, A. et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog,* 1.8: 9.

Ruttkay, Z. (1997). Composing Mozart variations with dice. *Teaching Statistics,* 19.1: 18-19.

Shaw, P., Uszkoreit, J., and Vaswani, A. (2018). Self-attention with relative position representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2

Wu, S., and Yang Y. (2020). The jazz transformer on the front line: Exploring the shortcomings of ai-composed music through quantitative measures. *arXiv preprint arXiv:2008.01307.*

Representing and Modeling Cultural Relevance in Corpora for Historical Analysis

Schroeter, Julian
julian.schroeter@uni-wuerzburg.de
University of Wuerzburg, Germany

Working with large corpora is one of the strengths of computational literary studies (CLS). Collecting metadata is the essential step for interpreting analytical results. So far, large corpora focus on non-historical metadata such as gender or general genres (Piper 2018, Jockers 2013). Historical metadata, if any, are collected as epoch or periods of publication, and with a growing interest also in the historicity of genres (cf. Underwood 2019, Schröter 2019). In CLS, there is, however, a growing interest in more specific socio-historical changes. Pursuing this interest necessitates collecting contextual metadata (cf. Riddell/Bassett 2020). One of the most important dimensions of a social history of literature is that of cultural relevance. So far, corpora tend to have several methodological weaknesses: Firstly, as most studies do not pay attention to the aspect of relevance, most corpora give all texts the same weight. In studies that have an interest in prestige, it is common practice to model these categories in a non-historical fashion (Algee-Hewitt/McGurl 2015) or from the perspective of present readers (Koolen et al. 2020). Secondly, most corpora have a bias towards highly canonical texts because canonical texts are already digitized whereas the ›big unread‹ awaits to be exploited. To investigate the historical change of cultural relevance, both weaknesses have to be overcome.

The central assumption behind the poster that shall be presented is that the forms of cultural relevance are much more diverse and that significant new historical insights can be gained by representing this diversity depending on specific research interests. Hence, the poster will have three areas that correspond to methodological steps, respectively:

1. It shows in a spreadsheet the current state of research.
2. It discusses the advantages and shortcomings of possible models of prestige based on visualizations starting from a specific interest in the dependency between media formats and genre semantics.
3. Using different forms of visualization, the poster demonstrates the impact of cultural relevance beyond prestige based on a genuine new corpus.

As this corpus includes more than 700 mostly forgotten novellas that were relevant in their contemporary contexts with a wide variety of historical metadata, it overcomes both weaknesses mentioned above. [1]

On the first issue (I), two aspects can be distinguished, the epistemic aspect of criteria for detecting prestige and different concepts of prestige. Table 1 outlines both aspects as two dimensions and provides a new view on prestige. On the poster, this table will be extended to the whole field of relevant research.

concept of relevance	epistemic dimension	form	source
›success‹ or ›popularity‹	sales	binary: Bestseller/no bestseller	Algee-Hewitt/ McGurl (2015)
›literary esteem‹	established lists of canonical works (»found lists«)	binary: in/not in corpus	Algee-Hewitt/ McGurl (2015)
›elitist prestige‹	being reviewed at least once	binary	Underwood/ Sellars (2016)
›elitist prestige‹	interviewing experts (»made lists«)	ranked positions	Algee-Hewitt et al. (2015)
Literary canon	Expert decision	ordinal scale of canon status: low, medium, high	ELTeC-Korpus [2]

Table 1: Conceptual and epistemic dimensions of modeling prestige

On the second issue (II), models for representing prestige are presented based on two conceptual dimensions: On the first dimension of context-sensitivity, there are two types of models: The context-insensitive model directly links a specific aspect of prestige to each literary work. The context-sensitive type of model, in contrast, links a specific aspect of prestige relative to a historical situation where the respective aspect of prestige was assigned to the work. The latter model carries the advantage that it facilitates investigating the assignment of prestige as a contingent social practice rather than interpreting prestige as eternal literary value.

The second dimension is that of a diversity of prestige. Three options are relevant: It is common practice to use only one concept and one epistemic aspect from table 1 as the ›best‹ proxy to cultural relevance. Another practice would be that of merging different epistemic dimensions into one encompassing idea of relevance. This strategy has the advantage that it takes into account different dimensions and provides only one resulting ratio that can be used in downstream tasks as one singular feature. Several

shortcomings have to be considered, such as the issue of scaling different epistemic aspects. Finally, there is an option of modeling a multi-dimensional space of relevance that preserves all different aspects.

On the third issue (III), the historical situation becomes more complex by the fact that historical relevance is not only a matter of prestige but also a matter of circulation. Three more dimensions have to be taken into account:

1. The degree of circulation in public libraries and reader circles.
2. The degree of supra-regional circulation
3. The average circulation volume of media types.

The poster will visualize the available data for the three types of circulation and dominance. Figure 1 provides a visualization of average circulation volume.

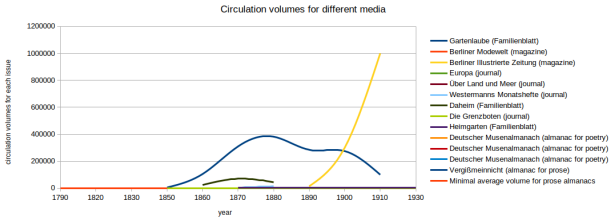


Figure 1:
Circulation volumes for different media

The poster is intended to encourage scholars who are interested in historical insights into literary cultures to discuss the ›next generation‹ of context-sensitive metadata acquisition and representation. Based on a specific historiographical interest in the dependency between media and genre, it aims to provide data-driven ground and methodological reasoning for discussing different options of generating context-sensitive corpora and to provide an argument for the desirability of such corpora in literary studies.

Bibliography

Algee-Hewitt, M. and McGurl, M. (2015). Between Canon and Corpus: Six Perspectives on 20th Century Novels, Literary Lab. litlab.stanford.edu/LiteraryLabPamphlet8.

Jockers, M. L. (2013). Macroanalysis: Digital Methods and Literary History. Urbana: University of Illinois Press.

Koolen, C., van Dalen-Oskam, K., van Cranenburgh, A., and Nagelhout, E. (2020). Literary quality in the eye of the Dutch reader: The National Reader Survey, Poetics 79: 1–13.

Piper, Andrew (2018). Enumerations. Data and Literary Study. Chicago ; London: The University of Chicago Press.

Riddell, A. and Troy, J. B. (2020) The Class of 1838: A Social History of the First Victorian Novelists, Mémoires du livre / Studies in Book Culture 11,2: 1–37.

Schröter, J. (2019). Gattungsgeschichte und ihr Gattungsbegriff am Beispiel der Novellen«, Journal of Literary Theory 13,2: 227–57.

Underwood, T. (2019). Distant horizons. Chicago, London: The University of Chicago Press.

Underwood, T. and Sellars J. (2016). The Longue Durée of Literary Prestige, Modern Language Quarterly 77,3: 321–44.

Notes

1. The corpus is described and can be explored in https://github.com/julianschroeter/19CproseCorpus.
2. Cf. the cost-action project, URL: https://www.distant-reading.net/(accessed December 8, 2022).

Developing a Comprehensive Application for Digital Transformation of Historical Materials

Shibutani, Ayako
ashibutani@hi.u-tokyo.ac.jp
Historiographical Institute, the University of Tokyo, Japan

Nakamura, Satoru
nakamura@hi.u-tokyo.ac.jp
Historiographical Institute, the University of Tokyo, Japan

Yamada, Taizo
t_yamada@hi.u-tokyo.ac.jp
Historiographical Institute, the University of Tokyo, Japan

Yanbe, Koki
yanbe@hi.u-tokyo.ac.jp
Historiographical Institute, the University of Tokyo, Japan

Scientific studies of historical materials have developed over the last decade using sophisticated analytical approaches (Shibutani, 2022). In particular, USB digital cameras for microscopy and megapixel camera lenses have

been upgraded, and researchers can easily obtain distortion-free and high-definition images (Shibutani & Goto, 2020). Most scientific data, including microscopic images, are now available in the digital format. However, preliminary analysis data cannot be shared for comparison processes because these analyses are conducted by 'specialised' researchers (Shibutani, 2020; Shibutani et al., 2021). Image data are usually rich data files because they present various parameters in a multidimensional space and are acquired using complex microscopy instruments. The real benefit of the easy sharing and reuse of digital data is that they aid data provenance and reproducibility of results. International standardisation of observational data modelling approaches is needed. With more openly accessible resource data, researchers can enhance and accelerate scientific advances in history. This study aims to develop the use of open science in history using an image data management tool. This aim is accomplished through the following three objectives: 1) conduct data provenance and lineage in history; 2) assess the impact of this new application in the science of historical materials; and 3) formulate recommendations to researchers on the appropriate strategy to promote reproducibility.

Within the analyst community, the image data management process is challenging, time-consuming, and difficult to scale. Researchers and analysts are seeking ways to manage image data effortlessly, quickly, and in higher quality. Our tool, 'classification and annotation for image data' (caid), is a new comprehensive application for image data management. Its main function is to manage research data provenance and lineage of multi-layered information of historical materials (Figure 1). The users of caid can preserve any resource data and update content easily. The application can be operated both online and offline.

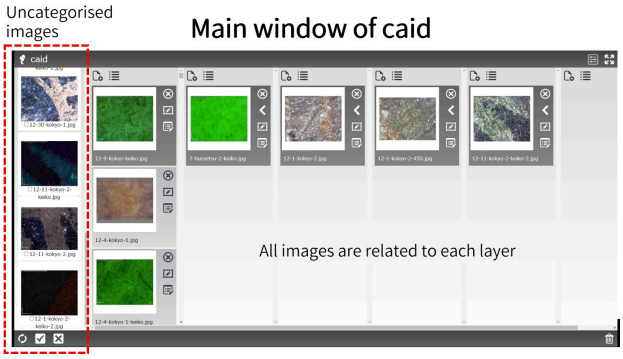


Figure 1.
Main window of caid

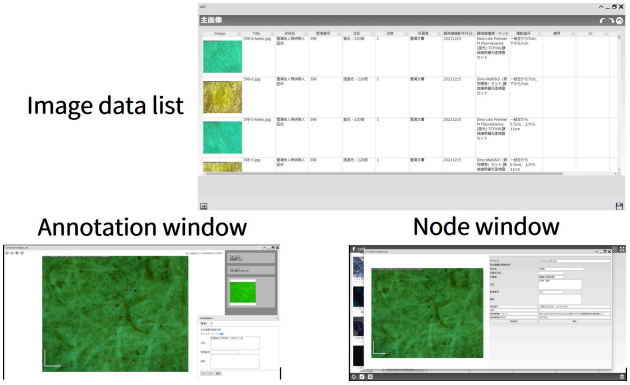


Figure 2.
Image data list and annotation and node windows of caid

Another function caid provides is fast and easy image annotation (Figure 2). Images require multiple labels for specifying contextual similarities during image retrieval. The caid provides an easy and comfortable method to label and edit metadata. Users can, thus, easily describe their notes during and after each survey. The labelling data will be connected to our institution's image digitalisation management system, which follows the Reference Model for an Open Archival Information System (OAIS). In addition, all input forms can be customised via content configuration and load specifications. Full descriptions of the surveyed materials can be added or replaced by the user during metadata collection. The temporal usability of our tool in a real survey shows the relevance of such technology in the field.

In scientific analyses, reproducibility, comparison, and sometimes integration of results are required. All the metadata in the caid can be saved and reloaded by the user for reuse and adaptation. This functionality allows the current survey data to easily compare with analytical results of/from other historical materials. The addition of different interpretations and annotations by different users to the same image data leads to conflicts, but their comparison on the caid refers to appropriate datasets. This comparison environment can support the study of the science of historical materials. Our presentation focuses on some cases showcasing the analytical ability of this application.

The caid seeks to improve research process by linking it to information infrastructure. It can solve the technological and sociological challenges that have limited open access to resource data worldwide. The explosion of artificial intelligence technology has made breakthroughs in image processing of scientific analyses (e.g. Haenlein & Kaplan, 2019; Savadjiev, et al., 2018). We will examine the applications of this system in the future. In doing so, our application can accelerate the digital transformation (DX) of historical materials.

Bibliography

Haenlein, M. and Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, 61 (4): 5-14.

Shibutani, A. (2020). Integrated studies of historical resources using archaeological and botanical methods: Perspectives of establishing "International Study of Historical Paper Materials". *Cultura Antiqua*, 72 (10): 82-89. (in Japanese)

Shibutani, A. (2022). Scientific study advancements: Analysing Japanese historical materials using archaeobotany and digital humanities. *Academia Letters, Article 4628*. https://doi.org/10.20935/al4628

Shibutani, A. and Goto, M. (2020). How Do Research Data Develop? International Standardisation of Scientific Data in Historical Studies. *Digital Humanities 2020: Conference Abstracts*. Online, July 2020.

Shibutani, A., Nomura T., Takashima, A., Masashi A., Yamada, T. (2021). Component Analysis of Historical Paper Materials at the Matsunoo Taisha Shrine Using Archaeological and Botanical Methodologies. *Tokyo Daigaku Shiryo Hensan-jo Kenkyu Kiyo*, 31: 59-74. (in Japanese)

Un nouveau partenariat sur les éditions critiques en contexte numérique

Sinatra, Michael
michaelesinatra@gmail.com
Université de Montréal, Canada

Vitali-Rosati, Marcello
marcello.vitali.rosati@umontreal.ca
Université de Montréal, Canada

Chateau-Dutier, Emmanuel
emmanuel.chateau.dutier@umontreal.ca
Université de Montréal, Canada

Nous souhaitons présenter dans un poster les enjeux au cœur du projet « **Nouvelles réflexions sur les éditions critiques en contexte numérique** » qui réunit des partenaires, des praticiens et des théoriciens provenant de quatre pays, impliqués dans plusieurs projets concernés au premier chef par les outils et la culture numériques, se positionne au centre des travaux de réflexion sur ces nouvelles formes d'environnements de recherche. Ce projet se base sur les travaux du Groupe de recherche sur les éditions critiques en contexte numérique (GREN). Après plusieurs années d'activités, les membres du groupe ont souhaité maximiser son expertise en collaborant de façon étroite avec quatre centres de recherche internationaux et 13 chercheurs universitaires américains, canadiens, allemand et français pour qui la question des éditions critiques est au cœur de leurs travaux, ainsi qu'avec une maison d'édition ouverte à expérimenter sur des nouvelles formes de publications savantes.

Subventionné par le Conseil de Recherche en Sciences Humaines du Canada de 2021 à 2024 dans le cadre du programme de développement de partenariat, le thème central de notre projet de développement de partenariat consiste à cerner les changements entraînés par le numérique dans le domaine de la production et de la diffusion du savoir dans les humanités. Plus précisément, les membres de notre projet étudient comment le numérique provoque un réagencement des dispositifs de production et de diffusion de la connaissance et des contenus en littérature, en communication, en histoire de l'art. Pour ce faire, ils s'appuient sur une expertise de haut niveau dans les sciences de l'information et l'informatique. Cette conjonction opportune de chercheurs et de domaines de spécialités nous a permis de jeter les bases d'une analyse et d'une appréhension à la fois pratique et théorique d'un objet d'étude crucial dans nos domaines respectifs, soit l'édition critique (au sens large) et ce qu'il en advient à l'ère du numérique (ici, édition critique étant entendue comme la production savante d'un document accompagnée d'annotations, de commentaires et autres matériaux jugés opportuns par l'éditeur). La nature même des éditions critiques a subi des changements majeurs avec l'explosion des données et la transformation des savoirs, et ce, tant sur le plan de la quantité d'information disponible que de la manière de la représenter. L'étude de ces changements et l'exploration concrète de nouvelles formes d'éditions critiques via l'implémentation d'outils informatiques sont au cœur de cette recherche ambitieuse et innovante.

Souhaitant encourager un dialogue entre théories et pratiques de l'éditorialisation, avec une série d'expérimentations basée sur les projets déjà en cours de nos chercheurs, notre projet de développement de partenariat compte réfléchir aux deux objectifs suivants :

1. Au plan des pratiques : dresser un état des lieux des expériences existantes en inventoriant les points forts et les points faibles des outils et des plateformes ; discuter de possibles stratégies de développement ; expérimenter l'utilisation d'outils de fouille et de visualisation ; intégrer ces outils dans des projets en cours et en développement ;

2. Au plan théorique : cerner l'impact du numérique sur le processus de production et de circulation du savoir : quels nouveaux modèles de lecture / écriture peut-on concevoir pour les éditions critiques du 21 e siècle? Quels dispositifs de validation des contenus, quels rapports entre chercheurs et communautés scientifiques?

Notre projet a été conçu de manière interinstitutionnelle afin de refléter la maturité des humanités numériques, avec une masse critique de chercheurs provenant d'institutions canadiennes mais aussi complémenté par des chercheurs reconnus aux États-Unis et en Europe. Les chercheurs impliqués dans notre équipe présentent un taux d'activité fort élevé dans le secteur, notre équipe étant composé de professeurs à différents stades de leurs carrières (nouveaux professeurs adjoints comme professeurs titulaires avec plus de vingt ans de carrière) qui ont également été impliqués dans de multiples collaborations. Nos chercheurs rassemblent des compétences essentielles au succès de notre initiative. L'appartenance disciplinaire des chercheurs et partenaires est distribuée au sein de départements de littérature, de communication, d'histoire de l'art, de philosophie et des sciences de l'information, et rassemble un ensemble d'expertises diversifiées et parfaitement adaptées à l'étude des éditions critiques en contexte numérique.

Notre poster mettra en avant le résultat des travaux de la première année.

Bibliography

Site du GREN: https://gren.openum.ca

HTR2CritEd: A Semi-Automatic Pipeline to Produce a Critical Digital Edition of Literary Texts with Multiple Witnesses out of Text Created through Handwritten Text Recognition

Stoekl Ben Ezra, Daniel
daniel.stoekl@ephe.psl.eu
EPHE, PSL, France; AOrOc UMR 8546

Lapin, Hayim
hlapin@umd.edu
University of Maryland, College Park, MD, USA

Brown-DeVost, Bronson
bronsonbdevost@gmail.com
EPHE, PSL, France; AOrOc UMR 8546; University of Goettingen, Germany

Jablonski, Pawel
pawel.jablonski@etu.ephe.psl.eu
EPHE, PSL, France; AOrOc UMR 8546

Database structures and export formats of Handwritten Text Recognition tools (e.g. Transkribus, Tesseract, eScriptorium) are usually based on a document layout hierarchy with regions/zones and lines. Interlinear or marginal additions to the main text are in separate zones and lines (Page XML, Alto XML) (Kahle et al. 2017, Stokes et al. 2021).

While this is less problematic for documentary texts (Chagué et al. 2021), it poses a problem for those working on critical editions of literary texts with multiple textual witnesses because any such edition presupposes a running text hierarchy (books, chapters, verses), where the interlinear and marginal additions need to be inserted at the right spots. This is a precondition to using text-alignment tools such as CollatEx (Dekkers et al. 2011).

We present a pipeline that permits to overcome this problem for Medieval Hebrew manuscripts in a semi-automatized fashion beginning with the discovery of insertion marks in the HTR process and leading to a critical edition in TEI:

1. We include a series of different insertion marks in the recognition training data for the HTR. Different insertion marks distinguish between interlinear and marginal additions (Stökl Ben Ezra et al. 2021).
2. Optimal matches of insertion marks with a) interlinear lines and b) marginal additions are calculated with the "Hungarian Algorithm" (Kuhn 1955). The results can be visualized via eScriptorium's API for image annotation (see fig. 1).
3. A. If there is already an e-text of a printed edition with an accepted text hierarchy, we use the Dicta synopsis-algorithm via an API (Brill et al. 2020) or alternatively a combination of global and local alignments of the Smith-Waterman (1981) and Needleman Wunsch (1970) algorithms to align the main text of the HTRed manuscript with the standard edition to calculate the places for the textual hierarchy markers. This needs to be manually verified subsequentially, especially if some of the markers for the text hierarchy should be in the interlinear or marginal additions.
   B. If there is no printed edition, the text hierarchy markers need to be inserted manually. This is usually

necessary only for one manuscript (if there is a manuscript that represents the complete text).
4. Based on the combination of 2 and 3, the first manuscript transcription in the HTR tool can now be converted from document hierarchy to text hierarchy TEI. If there was no printed edition (3B), the text hierarchy markers of this manuscript can be used in step 3A to automatically insert them.
5. The resulting data is submitted via json to an optimized Needleman-Wunsch algorithm, Collatex or another alignment tool (Brill et al. 2020) to automatically produce an alignment between the different witnesses. For error correction, Microsoft Excel can be used or the tool in step 7.
6. Text comparison in the alignment can serve to resolve most of the abbreviations.
7. The final result is fed into TEI-Publisher (Turska and Meier 2021). We hope to be able to integrate a tabular tool that allows to manually but ergonomically correct any misalignments of the automatic alignment process to produce the critical edition: https://editions.erabbinica.org/
8. The TEI-Publisher publication includes accessibility via DTS (Distributed Text Service).

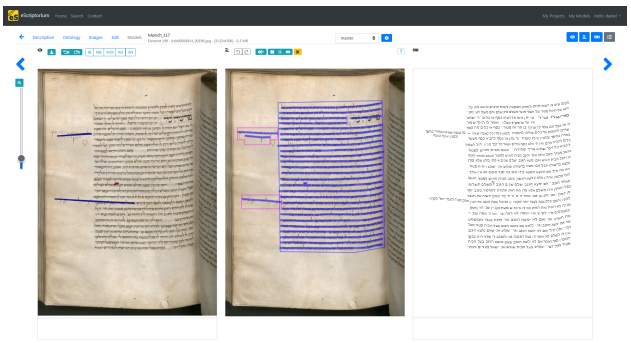


Fig. 1:
eScriptorium with 3 panels turned on: On the left, the image annotation panel with the triangles representing the links between marginal (blue) or interlinear (red) insertion spots and the first and last word of the insertion, the segmentation panel. In the center, the image annotation panel with the main text and the marginal and interlinear text lines. On the right, the manually corrected automatic transcription.

Bibliography

Almas, B., Clérice, T., Cayless, H., Jolivet, V., Liuzzo, P., Romanello, M., Robie, J., and Scott, I. (2021). Distributed Text Services (DTS): a Community-built API to Publish and Consume Text Collections as Linked Data (https://hal.archives-ouvertes.fr/hal-03183886)

Brill, O., Koppel, M., and Shmidman, A. (2020). FAST: Fast and Accurate Synoptic Texts. *Digital Scholarship Humanities* 35(2): 254-264.

Chagué, A. and Scheithauer, H. (2021). page2tei - LECTAUREP [Computer software]: https://github.com/lectaurep/page2tei

Dekker, R. H. and Middell, G. (2011). Computer-Supported Collation with CollateX: Managing Textual Variance in an Environment with Varying Requirements. *Supporting Digital Humanities* University of Copenhagen, Denmark. 17-18 November 2011.

Kahle, P., Colutto, S., Hackl, G., and Mühlberger, G. (2017). Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents," OST@ICDAR 2017: 19-24

Kuhn, H. (1955). "The Hungarian Method for the assignment problem," *Naval Research Logistics Quarterly*, 2: 83–97.

Needleman, S. and Wunsch, C. (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins," Journal of Molecular Biology 48 (3): 443–53.

Smith, T . and Waterman, M. (1981). "Identification of Common Molecular Subsequences." Journal of Molecular Biology 147 (1): 195–197.

Stokes, P., Stökl Ben Ezra, D., Kiessling, B., Tissot, R. and Gargem, E. (2021). "The eScriptorium VRE for Manuscript Cultures" in Claire Clivaz and Garrick V. Allen (eds), *Ancient Manuscripts and Virtual Research Environments*, special issue, *Classics@* 18 n.p.

Stökl Ben Ezra, D., Brown-DeVost, B., and Jablonski, P. (2021). "Exploiting Insertion Symbols for Marginal Additions in the Recognition Process to Establish Reading Order" *IWCP@ICDAR* 2021.

Turska, M., Meier, W. (2021). *TEI Publisher* http://teipublisher.com. (version 7, 2021)

Acceptable/Unacceptable/In-between Sentences in Japanese: An Experimental Study on Long-Distance Numeral Quantifiers

Suzuki, Kazunori
suzuki.kznr@gmail.com
Tokyo Institute of Technology

Hirano, Michiru
michirano10@gmail.com

Tokyo Institute of Technology

Yamamoto, Hilofumi
yamamoto.h.al@m.titech.ac.jp
Tokyo Institute of Technology

This study empirically examines the acceptability of sentences containing long-distance numeral quantifiers in Japanese in order to investigate the boundary between language competence and language use. In Japanese language, there is a grammatical phenomenon called long-distance numeral quantifiers (or floating numeral quantifiers, Miyagawa, 1989), in which a noun and the quantifier modifying the noun are grammatically acceptable even if they are not adjacent. When a noun and the numeral quantifier that modifies the noun are adjacent, it is grammatical and acceptable, but when other elements intervene between them, the acceptability may be maintained or it may fall. In other words, in sentences containing long-distance numeral quantifiers, the acceptability of a sentence depends not only on whether the noun and the numeral quantifier modifying the noun are adjacent or not, but also on the intervening factors between the noun and the quantifier even when they are remote.

In the present study, 30 adult native speakers of Japanese were asked to judge whether sentences containing numeral quantifiers in Japanese were acceptable or not. A total of 50 stimuli presented in the form of conversational sentences were prepared, and the participants were instructed to judge whether the underlined parts (sentences containing numeral quantifiers) in the conversational sentences were natural or not from "1 (unnatural)" to "4 (natural)", and to circle them with using a pencil. The types of the 50 sentences were: five types in which nouns and quantifiers were adjacent (15 sentences in total), five types in which nouns and quantifiers were not adjacent (25 sentences in total), and one type of distractor items (10 sentences in total). No time limit was set for the experiment, but it generally took about 20 minutes.

Prior to data analysis, test items for which "1 (unnatural)" or "2 (slightly unnatural)" were selected were replaced with "0", and test items for which 4 (natural) or "3 (slightly natural)" were selected were replaced with "1". Then, for all participants, the acceptance rate was produced for each type. Based on the acceptance rate for each type, we sorted the items from highest to lowest acceptance rate and created a ranking list.

The results of the data analysis revealed two points: (1) when the noun and the numeral quantifier modifying the noun were adjacent, the acceptance rate was high for all sentence types as shown in theoretical linguistic studies (e.g., Miyagawa, 1989; Miyagawa and Arikawa, 2007); (2) when the noun and the numeral quantifier modifying the noun were not adjacent, there were three types of acceptance rates: high, medium, and low. Although the medium acceptance type is analyzed as unacceptable in the field of theoretical linguistics research, the results of this study indicate the existence of a gray zone in terms of actual language use that is not completely unacceptable. That is to say, although the sentence structure or the word order is equivalent, the properties of the intervening elements between the noun and the numeral quantifier that modifies the noun are shown to affect the acceptability of the sentence. In order to elucidate human language functions, it is necessary to conduct an integrated study that includes language use (i.e., language performance).

Bibliography

Miyagawa, S. (1989). *Structure and Case Marking in Japanese (Syntax and Semantics 22)*. Academic Press.

Miyagawa, S., and Arikawa, K. (2007). Locality in syntax and floating numeral quantifiers. *Linguistic Inquiry*, 38(4), 645–670.

Characterizing playing style with speed deviation

Takahashi, Mai
klangvoll@gmail.com
The University of Tokyo, Japan

Kobayashi, Michikazu
kobayashi.michikazu@kochi-tech.ac.jp
Kochi University of Technology, Japan

Ohmukai, Ikki
i2k@l.u-tokyo.ac.jp
The University of Tokyo, Japan

Introduction

Recent performance researches have utilized computer software for analyzing recordings. Cock classified them as Distant Listening and Close Listening (Cook, 2014). To validate a transition of the performance style over a long span, Distant Listening tries to visualize quantitative commonalities and differences in a large number of recordings.

Past analyses of recordings have revealed that the performance style had changed around the 1920s. To distinguish performance styles before and after 1920s, the performance style before 1920s is often called “rhetorical performance” (Cook, 2014; Philip, 2004; Watanabe, 2001). Watanabe defined “rhetorical performance” as the performance style with drastic changes of the playing speed and oscillating rhythm (Watanabe, 2001). However, there is less objective and reproducible method established for analyzing such a change of the performance style and no decisive conclusion about it. In this work, by regarding performances as data, we try to reconstruct the performance theory cultivated in the past musicology based on the formal structure of performances. To do this, we propose a new objective method based on Distant Listening and discuss a change of the performance style from 1910s to now.

Target in analysis

We here focus on keyboard works by J. S. Bach; fugue part of *Chromatic Fantasy and Fugue* (BWV903), and prelude and fugue parts of *The Well-Tempered Clavier*, Vol. 1, No. 1 (BWV846). We analyze 82 types of recordings from 1910s to 2010s with various formats that has been historically changed. We pick up “playing speed” as a present analysis that hardly depends on kinds of recording format and instrument, recording environments.

Methods of analysis

Most of past works have analyzed a part of the whole recording (Cook, 2014; Leech-Wilkinson, 2015; Zhou, 2019). However, excerpting a part from the whole recording can cause an arbitrary conclusion and makes quantitative comparison among different recordings quite difficult. We here propose the method for quantitative analysis.

When analyzing recordings, as well as Cook, Leech-Wilkinson, and Zhou and Fabian, we use Sonic Visualiser developed by Centre for Digital Music at Queen Mary, University of London. We first open the recorded data (wav format), then make a beat data through Sonic Visualiser. We next calculate the performing speed at each beat by obtaining the time between one beat and the next beat. We finally calculate the speed deviation as an index of global change of the performing speed. The speed deviation is defined as the variation coefficient which is the standard deviation of the speed normalized by the average, enabling quantitative comparisons among different recordings. We calculate the speed deviation for all recordings for three works, and visualize the change of performance style.

Results

Figure 1 shows speed deviations for fugue part of *Chromatic Fantasy and Fugue*. Recordings denoted by blue circles have very large speed deviations and we can conclude them as “rhetorical performance” in nineteenth century. Recordings denoted by green circles also have large speed deviations and we define them as “quasi-rhetorical performance”. On the contrary, recordings denoted by red colors have very small speed deviations and we define them as “anti-rhetorical performance”.

Figure 2 shows speed deviations for prelude part of *The Well-Tempered Clavier*, Vol. 1, No. 1. We can see several recordings in 2000s regarded as “rhetorical performance”.

Figure 3 shows speed deviations for fugue part of *The Well-Tempered Clavier*, Vol. 1, No. 1.

What are commonly seen in above three works are as follows. Firstly, there existed “rhetorical performance” having large speed deviations before 1920s, which is consistent with previous researches (Watanabe, 2001; Leech-Wilkinson, 2015). “Interpretive editions” published before early twenty century (Bach, 1863; Bach, 1894) also have additional descriptions by editors regarded as “rhetorical performance”. Secondly, there are many recordings with “quasi-rhetorical performance” and is no recording with “anti-rhetorical performance” from 1930 to 1950. Thirdly, on the other hand, the number of recordings with "quasi-rhetorical performance" decreases and that with "anti-rhetorical performance" increases from 1960s to 1990s. Finally, recordings with “rhetorical performance” and “quasi-rhetorical performance” appear again in 2000s.

Conclusion

For three works analyzed in this paper, “rhetorical performance” was the main performance style before 1920s, and its vestige remained from 1930s to 1950s. From 1960s to 1990s, “rhetorical performance” declines once, and after 2000s, it has been appeared again. We can obtain this conclusion with objective and reproducible method, discussing a change of the performance style after 1930s. One of important conclusions is that characteristics of rhetorical performance with the large speed deviation has been remaining since 2000s. As a possible reason, we can assume that people have been able to share and easily get past recordings over the internet. Further researches are needed to certify this assumption for recordings not only of Bach but also other composers.

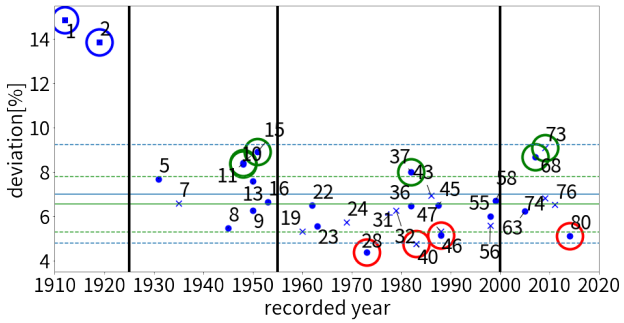


Figure 1.
Speed deviations for fugue part of Chromatic Fantasy and Fugue

Vertical and horizontal axes show the speed deviation and the recorded years, respectively. The number for each plot corresponds to the performer shown in Table 1. Circle, square, x, and plus symbols correspond to piano, piano-roll, cembalo, and clavichord, respectively. Horizontal solid blue and dashed blue lines show the mean value and the deviation from the mean value (standard deviation) of all speed deviations, respectively. All recordings with speed deviations large than the upper dashed blue line are denoted by blue circles. Solid green and dashed green lines show the mean value and deviation from the mean value of all speed deviations except for recordings with blue circles.

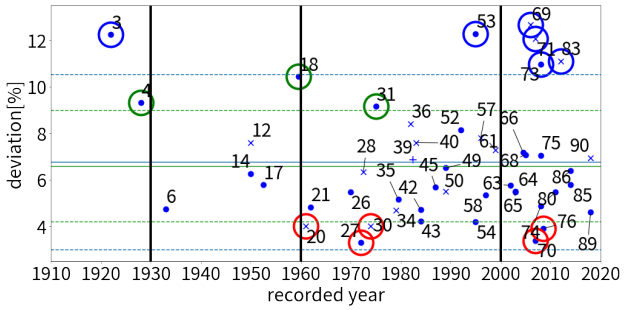


Figure 2.
Speed deviations for prelude part of The Well-Tempered Clavier*, Vol. 1, No. 1*

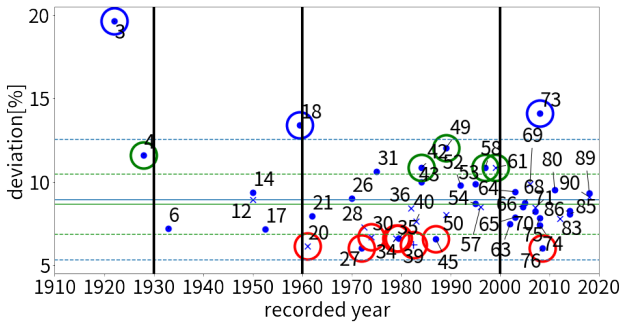


Figure 3.
Speed deviations for fugue part of The Well-Tempered Clavier*, Vol. 1, No. 1*

Nr.	year	performer	inst.	CFF	WTK	Nr.	year	performer	inst.	CFF	WTK
1	1912	F. Busoni	roll	○		42	1984	A. Schiff	pf		○
2	1912	W. Backhaus	roll	○		43	1987	T. Koopman	cem	○	
3	1922	F. Busoni	pf		○	44	1987	K. Jarrett	pf		○
4	1928	H. Cohen	pf		○	45	1987~88	I. Moravec	pf	○	
5	1931	E. Fischer	pf	○		46	1988	K. Lifschitz	pf	○	
6	1933	E. Fischer	pf		○	47	1988	S. Ross	cem	○	
7	1935	W. Landowska	cem	○		48	1989	J. Demus	pf		○
8	1945	C. Arrau	pf	○		49	1989	G. Wilson	cem		○
9	1948	E. Gilels	pf	○		50	1992	T. Sonoda	pf		○
10	1948	A. Schnabel	pf	○		51	1995	V. Afanassiev	pf		○
11	1948	M. Yudina	pf	○		52	1995	J. Jandó	pf		○
12	1949	W. Landowska	cem		○	53	1996	M. Suzuki	cem		○
13	1950	W. Gieseking	pf	○		54	1997	A. Hewitt	pf		○
14	1950	W. Gieseking	pf		○	55	1998	T. Sonoda	pf	○	
15	1951	E. Petri	pf	○		56	1998	M. Suzuki	cem	○	
16	1953	W. Kempff	pf	○		57	1999	L. Berben	cem		○
17	1953	R. Tureck	pf		○	58	1999	M. Koyama	pf	○	
18	1958~61	S. Feinberg	pf		○	59	2002	T. Fellner	pf		○
19	1960	H. Walcha	cem	○		60	2003	D. Barenboim	pf		○
20	1961	H. Walcha	cem		○	61	2003	K.-A. Kolly	pf		○
21	1962	G. Gould	pf		○	62	2004~05	V. Achkenazy	pf		○
22	1962	P. B.-Skoda	pf	○		63	2005	L. Fleisher	pf	○	
23	1963	L. Oborin	pf	○		64	2005	A. Vieru	pf		○
24	1969	K. Richter	cem	○		65	2006	R. Egarr	cem		○
25	1970	S. Richter	pf		○	66	2007	I. Nodaira	pf		○
26	1972	F. Gulda	pf		○	67	2007	M. Sone	cem		○
27	1972~73	G. Leonhardt	cem		○	68	2007	R. Woodward	pf	○	
28	1973	A. Brendel	pf	○		69	2008	A. Hewitt	pf		○
29	1974	H. Walcha	cem		○	70	2008	M. Stadtfeld	pf		○
30	1975	W. Kempff	pf		○	71	2008	R. Woodward	pf		○
31	1977~78	F. Gulda	cla	○		72	2009~09	M. Pollini	pf		○
32	1979	G. Leonhardt	cem	○		73	2009	C. Arita	cem	○	
33	1979	E. Picht-Axenfel	cem		○	74	2009	L. Berben	cem	○	
34	1979~80	M. Horszowski	pf		○	75	2011	A. Schiff	pf		○
35	1982	T. Koopman	cem		○	76	2011	Y. Watanabe	cem	○	
36	1982	T. Nikolayeva	pf	○		77	2012	A. Newman	cem		○
37	1982	A. Schiff	pf	○		78	2014	P. -L. Aimard	pf		○
38	1982~83	D. Chorzempa	cla		○	79	2014	I. Mejoueva	pf		○
39	1983	K. Gilbert	cem		○	80	2014	M. Stadtfeld	pf	○	
40	1983	E. Picht-Axenfel	cem	○		81	2018	C. Herrmann	pf		○
41	1984	T. Nikolayeva	pf		○	82	2018~19	T. Pinnock	cem		○

Table 1.
Details of recordings used in this paper

Bibliography

Cook, N. (2014). Beyond the Score: Oxford University Press.

Philip, R. (2004). Performing Music in the Age of Recording: Yale University Press.

Watanabe, H. (2001). Preface to Performance History (Japanese): Shunju.

Leech-Wilkinson, D. (2015). Cortot's Berceuse. Music Analysis, 34 (3): pp. 335-363.

Zhou, D. Q. and Fabian, D. (2019). A Three-Dimensional Model for Evaluating Individual Differences in Tempo and Tempo Variation in Musical Performance: Musicae Scientiae. doi: https://doi.org/10.1177/1029864919873124.

Bach, J. S. (1863). Ausgabe Hans von Bülow. Chromatische Phantasie: Ed. Bote & G. Bock.

Bach, J. S. (1894). Das wohltemperierte Klavier von Ferruccio Busoni: Breitkopf & Härtel.

Extracting clichés: Typify slanderous expressions against the confessions in the #MeToo movement

Takedomi, Yuka
yuka_takedomi@nii.ac.jp
National Institute of Informatics, Japan

Suda, Towa
sudatowa@nii.ac.jp
National Institute of Informatics, Japan

Kurita, Kazuhiro
kurita@nii.ac.jp
National Institute of Informatics, Japan

Kobayashi, Ryota
r-koba@edu.k.u-tokyo.ac.jp
The University of Tokyo/ JST PRESTO,

Matsuda, Tomohiro
tm0407@nii.ac.jp
National Institute of Informatics, Japan

Uno, Takeaki
uno@nii.ac.jp
National Institute of Informatics, Japan

The #Metoo movement, which burgeoned on social networking sites in October 2017, saw a series of private confessions and a large number of people experiencing the massive social movement. This movement became a milestone of the possibilities of transnational solidarity and had an unparalleled and powerful impact on society. In contrast, it has provoked a huge backlash, with social networking sites becoming the place of a wide variety of slanderous exchanges; some hurled abuse, others criticized that the confessions seemed unreliable and undermined their value.

The offensive expressions seriously impact the persons to whom they are directly addressed and those around them. People shrank at outrageous attacks, their dignity is violated, and they are often forced to be silent. However, if we can confirm any patterns in the abusive expressions and actions of others, which we currently perceive only as absurdity, and if we can make sure that these are clichés, our fear and psychological damage may be alleviated. Furthermore, if we wish to dispirit this kind of slander, we must begin by understanding the mechanisms of slander. Then how do we understand the apparent variety of slanderous expressions that have appeared on social networking sites in the #MeToo movement?

The appearance of slanderous expressions is diverse. However, there seem to be some typical psychological mechanisms and styles of expression that lead to those writing. Kate Manne, for example, uses the methods of analytic philosophy to examine the logic of misogyny and present a typology of the system (Manne, 2017). She argues that excessive bashing occurs when women visibly resist or violate social norms. In her view, these attacks are "not a matter of the psychology of individuals," but by the collective surveillance of women and the punishment of those who do not comply. She further distinguishes misogyny as a "law enforcement" branch; a combative system that attacks violators of the patriarchal order, and sexism as a "justificatory" branch of a patriarchal order; a theoretical system that justifies and theorizes patriarchal social norms and gender roles.

In this research, we aim to understand the typology theory obtained in the field of humanities, as exemplified above, from the quantitative point of view in the case of the #Metoo movement on Twitter. That is, we examine correspondences between qualitative theory and quantitative results from the Twitter data.

First, we collect social media posts (tweets) on Twitter about the following cases that have been popular in the #Metoo movement in Japan: 1) A case of Shiori Ito, who later became the symbol of Japan's #MeToo movement; she held a press conference in May 2017 accusing the journalist and published the book, 2) #KuToo movement; Yumi Ishikawa started a campaign to outlaw corporate practices that force women to wear high heels as sexual discrimination or harassment, 3) Flower Demo; The acquittals in four sexual assault cases in March 2019 triggered a nationwide grassroots movement.

Next, we extract slanderous words by discovering topics from the Twitter data. Specifically, we will apply topic models (e.g.,Latent Dirichlet Allocation, clustering algorithm (Blei, 2012)) and topic mining methods (e.g., Data Polishing (Uno, 2017; Hashimoto, 2021.)) for automatically finding the topics. By referring to the qualitative study of humanities, we typify and interpret the micro topics. In this way, we add objective explanations with an exhaustive approach to the concepts of the qualitative approach. At the same time, we apply the concepts from the qualitative approach to make sense of the typologies that are difficult to understand by using the exhaustive approach. In this way, we aim to fill in the

methodological gaps between the two approaches and to get a better overall picture of slander.

Bibliography

Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55, 77-84.

Hashimoto, T., Shepard D., Kuboyama, T., Shin, K., Kobayashi, R., and Uno T. (2021). Analyzing Temporal Patterns of Topic Diversity using Graph Clustering, The Journal of Supercomputing, 77, 4375-4388.

Manne, K. (2017). Down Girl: The Logic of Misogyny. Oxford University Press.

Uno, T., Maegawa, H., Nakahara, T., Hamuro, Y., Yoshinaka, R., and Tatsuta, M. (2017). Micro-clustering by Data Polishing. IEEE BigData 2017, 1012-1018.

Automatic matching method of historical event text with its corresponding thematic maps developed for the application of the ShiJi Spatio-Temporal Information Platform

Tsai, Jung-Yi
jungyitsai@gmail.com
Center for GIS, Research Center for Humanities and Social Sciences, Academia Sinica, Taiwan

Pai, Pi-Ling
lingpai@gate.sinica.edu.tw
Center for GIS, Research Center for Humanities and Social Sciences, Academia Sinica, Taiwan

Liao, Hsiung-Ming
veevee@gate.sinica.edu.tw
Center for GIS, Research Center for Humanities and Social Sciences, Academia Sinica, Taiwan

Chen, You-Jun
naomichen.yj@gmail.com
Center for GIS, Research Center for Humanities and Social Sciences, Academia Sinica, Taiwan; Department of Mathematics, University of California, Los Angeles, CA, USA

Tsai, Richard Tzong-Han
thtsai@g.ncu.edu.tw
Center for GIS, Research Center for Humanities and Social Sciences, Academia Sinica, Taiwan; Department of Computer Science and Information Engineering, National Central University, Taiwan

Fan, I-Chun
mhfanbbc@ccvax.sinica.edu.tw
Institute of History and Philology, Academia Sinica, Taiwan

The powerful data integration function of the spatio-temporal information technology has gradually been affirmed for its importance in the historical research and application of digital humanities under the establishment of a large number of digital historical text archives. Since spatio-temporal information is an important attribute of historical events, and historical maps are important carriers for presenting historical events, after extracting thematic events from historical texts, various thematic event maps can be generated based on the place name database and basic digital historical maps of Chinese Civilization in Time and Space (CCTS) (Academia Sinica, 2002). In addition, we can construct a thematic spatio-temporal platform to integrate and present event text with its corresponding maps.

The ShiJi Spatio-Temporal Information Platform is an integrated system of thematic historical maps and texts to present the historical events recorded in Records of the Grand Historian (Chinese name ShiJi) (Jung-Yi Tsai et al., 2021). The original data on the platform is mainly based on the dataset compiled by Historian Professor Panqing Xu, with a total of 1,260 historical events and 360 thematic maps (Xu, P. -Q., 2010). There is a one-to-many relationship between the text and the maps, and after the initial manual comparison, there are still about 760 historical events that cannot be clearly related to the thematic maps. Therefore, we designed a set of preprocessing procedures and algorithms for automatically sorting applicable maps to match historical events with the thematic maps.

There are two parts to preprocessing. First, we define the spatial scope of the historical event using the coordinates of place names in the event text, and evaluate the proportion of the place name coordinates of the historical event covered in the thematic maps, so as to quickly filter out irrelevant maps. The spatial overlap ratio of the event and the map is set to 0.8, that is, the thematic map where more than 80% of the place names of an event is located will be considered valid. Another preprocessing is to use the OCR tool (Rakpong Kittinaradorn, 2020) to extract the text annotations in the thematic maps for the subsequent matching algorithms.

In the automatic map matching algorithms designed in this research, the first step is to extract the place names in the text and the map based on the CCTS place name database; the second is to convert the place names in the text and the map into TF-IDF vector (Kim and Gil, 2019), and then calculate the cosine similarity to find the maps that overlap with the main location of the historical event. The third step is to convert the place names in the text and the map into one-hot vector, and then calculate the cosine similarity of the place name distribution, so as to improve the appearance of the place names appearing on the map; finally, we integrate the cosine similarity of the place names from the map and the text to sort the applicable maps.

From the dataset processed above, we select some examples with more spatial attributes to perform mean reciprocal rank (MRR) (Valcarce et al., 2020) experimental calculation, such as the "Appointing Pei Gong to Attack the West" in the Battle of Julu (207 BC), and the "Xiang Yu Marching to Xi" in the Hongmen Banquet (206 BC). The MRR calculated in this way is close to 80%. Through experiments, we also discovered some interesting spatio-temporal characteristics of historical events. For example, for the automatic map matching algorithms of the event "Appointing Pei Gong to Attack the West"" in the Battle of Julu, thematic maps with the same attack direction and route were found.

The automatic matching framework between the historical event text and the corresponding thematic map developed in this research has been implemented in ShiJi. In the future, it is expected to be further applied to the automatic linking of various historical texts and historical maps, to continuously improve the structure of this research, and to explore related research topics.

Bibliography

Academia Sinica (2002). *Chinese Civilization in Time and Space (CCTS)*. https://ctext.org/static/shanghai2018/liaohsiungming-geohumanities.pptx.

Jung-Yi Tsai, Pi-Ling Pai, Hsiung-Ming Liao, You-Jun Chen, Richard Tzong-Han Tsai, and I-Chun Fan (2021). Construction of ShiJi Spatiotemporal Information Platform on the Framework of Research-oriented Knowledge Bases. *JADH 2021*.

Kim, S.-W. and Gil, J.-M. (2019). Research paper classification systems based on TF-IDF and LDA schemes. *Human-Centric Computing and Information Sciences*, **9**(1). Springer: 1–21.

Rakpong Kittinaradorn (2020). *EasyOCR*. https://github.com/JaidedAI/EasyOCR.

Valcarce, D., Bellogín, A., Parapar, J. and Castells, P. (2020). Assessing ranking metrics in top-N recommendation. *Information Retrieval Journal*, **23**(4). Springer: 411–48.

Xu, P. -Q. (2010). *Atlas of ShiJi*. Beijing: Seismological Press.

Using Automated Textual Analysis to Study Concepts of Identity and Difference in First-Person Narrative

Wang, Yadi
yadiw@uw.edu
University of Washington

Cole, Camille Lyans
camillecole@gmail.com
University of Cambridge

Fields, Sam E.
fields8@uw.edu
University of Washington

Saelid, Daniel P.
saeliddp@uw.edu
University of Washington

Chen, Annie T.
atchen@uw.edu
University of Washington

Distant reading involves the use of computational methods to visualize and study text. Distant reading methods have been used with a variety of source texts and genres, including novels, biographies, tales, and news collections (Jänicke et al., 2015) to explore a range of research questions. However, researchers have also raised concerns about distant reading methods oversimplifying complexity (Ascari, 2014) and advocated for approaches combining close and distant reading to allow broader and more comprehensive analysis (Ascari, 2014; Drouin, 2014).

In this poster, we explore the affordances and pitfalls of one such distant reading method, word co-occurrence analysis, for the exploration of identity and difference in first-person writing, using two source texts which both appear in the format of a diary and yet are fundamentally different – the diary novel and the historical diary.

On the one hand, we consider Yone Noguchi's *The American Diary of a Japanese Girl*, the fictional diary of a

Japanese immigrant in the early twentieth-century United States. As the young heroine, Morning Glory, enjoys her transcontinental adventure by interacting with people with different gender, racial and ethnic roles, we are drawn into her playful exploration of these concepts and her own identity.

On the other hand, we study the historical diary of Joseph Mathia Svoboda, a purser on a British steamship in nineteenth century Iraq. While a large portion of Joseph Svoboda's diary accounts for his daily routine as a pursuer, due to the multiculturalism of the region, he too is embedded in a context in which people of different backgrounds interact, and his writing in turn affords us a view of perceptions of gender, race, and ethnicity, both his own and others.

Therefore, even though one is fictional and the other is historical, both diaries are rich in depictions of social interactions in the context of everyday life, from the perspective of a person/character who intermixes multiple languages in everyday writing. While the two works differ in terms of the author/character's identity and positionality within the society being described (immigrant vs local), both allow us to observe perceptions of identity and difference.

In this poster, we propose to employ automated textual analysis and visualization methods to facilitate exploration of concepts of identity and difference within these two texts. In particular, we plan to render visualizations of common word co-occurrences that relate to gender and race/ethnicity in each of the two works, and reflect on the utility of this method given the language/genre of each text. Can word co-occurrence analysis be used as a distant reading technique to help us examine representations of gender, race/ethnicity in literary and historical works, to provide new insights on social and cultural contexts in which they were written, and if so, how?

Because of the differences in genre between the two texts, the poster does not compare them directly. Rather, it reads them in parallel to understand what kind of insight word co-occurrence analysis can offer into concepts of identity and difference in first-person narratives. In addition, we reflect on the benefits and shortfalls of text analysis and visualization methods in addressing these questions (Chen and Cole, 2021).

Bibliography

Ascari, M. (2014). The Dangers of Distant Reading: Reassessing Moretti's Approach to Literary Genres. *Genre*, **47** (1): 1–19 doi:10.1215/00166928-2392348.

Chen, A. T. and Cole, C. L. (2021). Reflexivity in Issues of Scale and Representation in a Digital Humanities Project. *ArXiv:2109.14184 [Cs]* http://arxiv.org/abs/2109.14184 (accessed 9 December 2021).

Drouin, J. (2014). Close- and Distant-Reading Modernism: Network Analysis, Text Mining, and Teaching the Little Review. *The Journal of Modern Periodical Studies*, **5** (1). Penn State University Press: 110–35 doi:10.5325/jmodeperistud.5.1.0110.

Jänicke, S., Franzini, G., Cheema, M. F. and Scheuermann, G. (2015). On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. The Eurographics Association, p. 21 doi:10.2312/eurovisstar.20151113.

Issues on Text Encoding of an East Asian Literature

Wang, Yifan
747.neutron@gmail.com
Graduate school of the University of Tokyo / International Institute for Digital Humanities

Nagasaki, Kiyonori
nagasaki@dhii.jp
International Institute for Digital Humanities

Shimoda, Masahiro
shimoda@l.u-tokyo.ac.jp
Graduate school of the University of Tokyo

Despite the existence of huge amount of textual material, TEI has not been as widespread in East Asia as in the West. The reason for this may be due to socio-cultural differences, but also probably to the fact that East Asian characters were difficult to use on computers in the past, thus there was little incentive to address text encoding. However, as the characters have become somewhat easier to use, the environment is now ready for scholars to work on text encoding. This presentation will describe our attempts to encode an East Asian literature under the situation.

As a part of effort to digitalize the scholarly standard text of Chinese buddhist canon, *Taisho Tripitaka* 大正新脩大藏經, the authors are working on the markup of *Xu Yiqiejing Yinyi* 續一切經音義 (Xilin, 1928), which has a non-trivial structure. The text is a kind of dictionary which glosses characters, words, and their pronunciation occurred in Buddhist scriptures. As the headwords are sequenced by the order of occurrence in the scriptures, it provides at least two pieces of interesting information, that is: firstly, the glosses usually explain them in the

contexts of each text; and secondly, it preserves more traditional forms of the characters and the words rather than what in ordinary scriptures referred by the dictionary. Since the editor of the dictionary consulted or annotated against various manuscript texts of the Buddhist scriptures available in the 8 th century before woodcut printing became popular in East Asia, it often preserves the older forms than the texts we currently know, which has been changed through the historical inheritance. The feature makes it useful in reconstruction of the original texts and the old form of the characters. However, since many characters cited by the dictionary are marked as non-standard at that time, those characters are sometimes difficult to gain support for encoding, in spite of the increasingly academic attitude of character encoding communities including Unicode Consortium, ISO/IEC JTC1/SC2/WG2, and IRG (Ideographic Research Group).

According to the situation, encoding of the dictionary requires not only the structure of a dictionary, but also linkage to the targeted scriptures. So, we aim to mark it up according to the following policy based on the Best Practices Level 3 (*Best Practices for TEI in Libraries*, n.d.).

1. The dictionary has a common style in East Asia, that a headword is glossed by following doubled lines (Fig. 1). It is necessary to encode not only the headword and the gloss, but also the typical style of the sub-lines with a line break. We currently adopt <seg> and <lb/> with @type=”wari” to describe it (Fig. 2) but see the necessity of unambiguous handling of sub-linear layout with additional attribute(s).

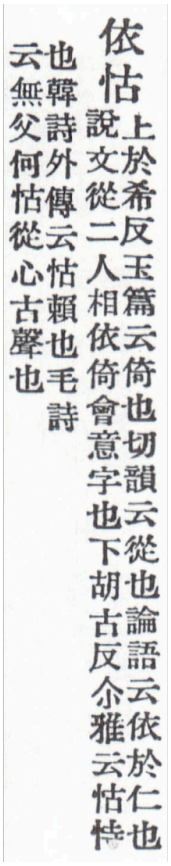


Fig. 1

```
172                 <lb n="T2129_,54,0935a07"/>依怙<seg type="wari">上於希反玉篇云倚也切韻云從也論語云依於仁也<lb
173 type="wari"/>說文從二人相依倚會意字也下胡古反尒雅云怙恃</seg>
173                 <lb n="T2129_,54,0935a08"/>  <seg type="wari">也韓詩外傳云怙賴也毛詩<lb type="wari"/>云無父何怙
173 從心古聲也</seg>
```

Fig. 2

2. The structure of the dictionary is related to both the physical form of itself and that of referenced scriptures. While the two divisions are sometimes overlapped, we constructed the basic hierarchy based on the latter and the former is marked up by <milestone/>.

3. As mentioned above, character encoding is an issue in this text. Although our proposal to encode several hundred characters occurred in this text have already accepted in the UCS (Universal coded Character Set), that is, Unicode, many remain unencoded due to not only the uncertainty such as erroneousness and blur, but procedural restrictions (Lu, 2022). In this case, gaiji module in TEI Guidelines (*Characters, glyphs, and Writing Modes*, n.d.) is useful. The following situations are seen in East Asian literature: a) appearance of characters not reducible to an existing UCS character are fairly common; b) with a high number of unencoded characters, they need to queue long for Unicode standardization and take progressive steps before completion; c) character sets with questionable Unicode compatibility are sometimes popular to cover missing

characters. The gaiji module is almost compatible with East Asian characters, but not enough for self-contained character description needed at this level. To solve the issue, we suggest the following extensions to the TEI Guidelines in the meantime.

- To add declaration of non-Unicode source to <localProp> as @scheme.
- To add minimum and maximum version numbers to <unicodeProp>, <unihanProp>, and <localProp> as @minVer and @maxVer for mutable properties.
- To add datable attributes (att.datable) to <mapping> for traceability of historical identities of a character in standards.

The extension provides a way to keep persistent reference of non-Unicode characters in TEI document.

Bibliography

Lu, Q. (ed). (2022). IRG Principles and Procedures (IRG PnP) Version 15. https://appsrv.cse.cuhk.edu.hk/~irg/irg/irg58/IRGN2515PnP15Confirmed.pdf (accessed 19 April 2022).

Xilin [希麟]. (1928). Xu Yiqiejing Yinyi [續一切經音義]. In Takakusu, J. and Watanabe K. [高楠順次郎・渡邊海旭] (eds), *Taishō Shinshū Daizōkyō* [大正新脩大藏經], vol. 54. Tokyo: Taisho Issai-Kyo Kanko Kwai, pp. 934–979.

Best Practices for TEI in Libraries. (n.d.). https://tei-c.org/extra/teiinlibraries/ (accessed 1 December 2021).

Characters, Glyphs, and Writing Modes. (n.d.). https://tei-c.org/Vault/P5/4.3.0/doc/tei-p5-doc/en/html/WD.html (accessed 10 December 2021).

Distance Reading Mary McCleod Bethune and the Black Fantastic

Williams, Seretha D.
seretha.williams@augusta.edu
Augusta University, United States of America

West-White, Clarissa
whitec@cookman.edu
Bethune-Cookman University, United States of America

Kizer, Ianna
whitec@cookman.edu
Bethune-Cookman University, United States of America

Dickey, Tierany
whitec@cookman.edu
Bethune-Cookman University, United States of America

Albury, Lauren
whitec@cookman.edu
Bethune-Cookman University, United States of America

Our digital humanities research project, "The Black Fantastic: Curated Vocabularies, Artifact Analysis, and Identification," proposes to locate uncategorized texts that share characteristics of known Black Fantastic texts and to build a workset, or a curated collection of digital volumes, that can serve as data used to understand the Black Fantastic as a genre or subgenre of literature. This research is supported by an award from the *Scholar-Curated Worksets for Analysis, Reuse & Dissemination* (SCWAReD) project and through the technical support of the Research Center of HathiTrust Digital Library. We use Richard Iton's term "Black Fantastic" to refer to the varied iterations of speculative inquiry and creative expression produced by Black people in the United States and the African Diaspora. The term "Afrofuturism," coined by Mark Dery, is contested in some scholarly and creative communities; thus, our use of "Black Fantastic" in our research acknowledges those debates and centers inclusive language.

While the project primarily focuses on our work using analytics and rooting out texts in the HathiTrust Digital Library, this poster focuses on a secondary research question and the challenges of incorporating undergraduate students into digital humanities research projects. Dr. West-White worked with students at Bethune-Cookman University, a historically black university, to digitize texts from the Mary McCleod Bethune collection in the Carl S. Swisher Library. Bethune, a visionary leader of education and civil rights, is frequently associated with W.E.B. DuBois, who in addition to his leadership as a sociologist, activist, and editor of Crisis, was a speculative fiction writer. His fiction, especially the story "The Comet" is generally recognized as an exemplar of early Black Fantastic writing. Our secondary research question, then, considers whether other early twentieth century social and political luminaries of the Black community were thinking and writing about race in ways that can be described as speculative or fantastic. By digitizing the correspondence, essays, and speeches of Mary McCleod Bethune, Dr.

West-White and her student assistants have attempted to identify textual evidence to

support the claim that Mary McCleod Bethune, like DuBois, was concerned with world building and imagining future spaces of actualized Black liberation. Bethune, the founder of Bethune-Cookman University, spoke and wrote

extensively about the future of Black people. In a 1926 speech at the convention of the National Association of Colored Women, Bethune told the audience, “I shall not review history because you are familiar with all that relates to our past and present in America. The present will be emphasized as a foundation of future prophecy” (Bethune 158). Bethune’s use of the phrase “future prophecy” here is the type of Black Fantastic diction this project seeks to uncover.

Using Brett Hirsch’s *Digital Humanities Pedagogy: Practices, Principles and Politics* and other scholarly books on the subject of pedagogy and DH, the faculty researchers developed a plan for including undergraduate researchers. The students hired for this project did not have digital humanities experience, so the faculty researchers trained students to digitize texts, clean up the corpuses, ensure the OCR worked and the text was searchable, conduct keyword searches, and notate patterns and characteristics that may indicate an instance of Black Fantastic concepts or ways of saying.

Additionally, students were asked to use GIS to map sites where collections of Mary McCleod Bethune writings are held outside of Bethune-Cookman University. Most of this work has been conducted virtually because students were not on campus due to the Covid-19 pandemic.

Digitization of selected materials in the Mary McCleod Bethune collection at

Bethune-Cookman University is complete. Our support partners at the Research Center of HathiTrust Digital Library are working with us to run TF-IDF on the materials from the Bethune-Cookman collection and the digital volumes contained in the HathiTrust collection. We will use keyword extraction to determine word frequency across the Bethune corpus, and we will analyze the keywords to determine whether words we associate with the verve of the Black Fantastic appear in the Bethune corpus.

Bibliography

Bethune, M.M. (2001). *Mary McLeod Bethune: Building a Better World: Essays and Selected Documents*. Bloomington: Indiana University Press.

Dery, M. (1993). Black to the Future: Interviews with Samuel R. Delany, Greg Tate, and Tricia Rose. *South Atlantic Quarterly*, 92.4: 735–778.

Hirsch, B.D. (2012). *Digital Humanities Pedagogy: Practices, Principles and Politics*. Cambridge: Open Book Publishers, http://books.openedition.org/obp/1605 (accessed 11 December 2021).

Iton, R. (2008). *In Search of the Black Fantastic*: *Politics and Popular Culture in the Post-Civil Rights Era.* Oxford: Oxford University Press.

Humanities Data Inquiry: A Community of Practice Exploring Data Issues in the Humanities and Heritage Research

Woods, Nathan
nathan.woods@uleth.ca
Humanities Innovation Lab, University of Lethbridge, Canada

Bordalejo, Barbara
barbara.bordalejo@uleth.ca
Humanities Innovation Lab, University of Lethbridge, Canada

O'Donnell, Daniel
daniel.odonnell@uleth.ca
Humanities Innovation Lab, University of Lethbridge, Canada

This poster presentation provides a portrait of the Humanities Data Inquiry (HDI), a new (2021) Social Sciences and Humanities Research Council (SSHRC)-funded project. The funding application for this project was entitled “Good things come in small packages: A grassroots Community of Practice for Open and FAIR humanities data practices.” HDI uses a research-informed Community of Practice (CoP) model to explore a diversity of issues involving the development, use and organization of humanities and cultural heritage (HCH) data, particularly when compared to ‘big data’ STEM and Open Science Frameworks. It outlines the basic problem associated with representational data in HCH and Open Science infrastructure frameworks and explains how HDI will address this through its CoP structure and research activities. Finally, the poster presents areas of future development and pathways for future involvement in the community of practice as well as the program’s expected outcomes.

The last decade has seen great advances in the development of infrastructure, tools, and principles for the collection, storage, discovery, and dissemination of research data (Borgman, 2015). Governments, funders, libraries and consortia, and private corporations have made large investments in what is rapidly becoming a robust Open and FAIR (Findable, Accessible, Interoperable, and Reusable) research data ecosystem.

This ecosystem, however, was built largely with the needs of "Big Data" Science, Technology, Engineering, and Mathematics (STEM) in mind. While Humanities

and Cultural Heritage (HCH) researchers and projects are encouraged to engage with this evolving research data ecosystem, the fit is often poor. Where Big Data STEM typically involves large datasets produced through experiment, observation, or analysis, "Small Data" HCH research often involves deep analysis and intensive curation of relatively small data sets — perhaps particularly when it comes to datasets (and data points) focused on the representation of cultural texts and objects: i.e., editions and exhibits (Oldman, 2021; Franzini et al., 2019).

Addressing this disjunction requires community engagement and input: thoughtful and bidirectional communication amongst a diversity of practicing humanities researchers and between researchers and the organizations responsible for creating and supporting the Open and FAIR ecosystem.

The poster outlines HDI's three main goals, designed to support this communication and community building:

- Discover: develop a forum for the development of best practice in the use of Open, FAIR, and CARE (Collective Benefit, Authority to Control, Responsibility, and Ethics)-compliant RDM practices among HCH researchers and projects;
- Demonstrate: Support the adaptation of Open, FAIR, and CARE RDM principles to Small Data HCH projects through the creation of a virtual "exolab" — a collaboration among leading data-centric research projects to share research problems in order to develop common approaches and practices at the level of the investigator;
- Mobilise: Promote the discovery, thoughtful adoption, and bidirectional development of Open, FAIR, and CARE RDM principles where appropriate by the broader HCH research community through peer-to-peer Workshops, Summer Schools, and training sessions.

While recent work has highlighted the importance of understanding scholarly information and work practices in the design of scholarly tools in the digital humanities (Antonijević and Stern-Cahoy, 2016; Lamb and Kling, 2003) in this work the goal is often framed as a question of how to include the voices of HCH researchers as user inputs, where problems are often scoped as questions of capacity and inclusion. By contrast, a unique feature of HDI is its use of ethnographic research (Borgman et al.; Koch, 2017; Ribes, 2014; Star, 1999; Wenger, 1999) and design (Poirier, 2017; Pyrko et al., 2017; Baker and Millerand, 2007; Ribes and Baker, 2007) to support community development, and enhance discovery and documentation through participatory agenda-setting activities. By bringing into conversation different sectors of the humanities and cultural heritage information ecosystems, the model will contribute to community-led design initiatives.

Bibliography

Antonijević, S. and Stern-Cahoy, E. (2016). Developing Research Tools via Voices from the Field. *DH+ LIB Special Issue*.

Baker, K. S. and Millerand, F. (2007). Scientific infrastructure design: Information environments and knowledge provinces. *Proceedings of the American Society for Information Science and Technology*, vol. 44. Wiley Online Library, pp. 1–9.

Borgman, C. L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, Massachusetts: The MIT Press

Borgman, C. L., Wofford, M. F., Golshan, M. S., Darch, P. T. and Scroggins, M. J. Collaborative Ethnography at Scale: Reflections on 20 years of Acquiring Global Data and Making Data Global.: 63.

Franzini, G., Terras, M. and Mahony, S. (2019). Digital editions of text: surveying user requirements in the digital humanities. *Journal on Computing and Cultural Heritage (JOCCH)*, **12**(1). ACM New York, NY, USA: 1–23.

Koch, G. (2017). The ethnography of infrastructures: Digital Humanities and Cultural Anthropology. *Cultural Heritage Infrastructures in Digital Humanities*. Routledge, pp. 63–81.

Lamb, R. and Kling, R. (2003). Reconceptualizing users as social actors in information systems research. *MIS Quarterly*. JSTOR: 197–236.

Oldman, D. (2021). Digital research, the legacy of form and structure and the ResearchSpace system. *Information and Knowledge Organisation in Digital Humanities*. Routledge, pp. 131–53.

Poirier, L. (2017). Devious design: Digital infrastructure challenges for experimental ethnography. *Design Issues*, **33**(2). The MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info …: 70–83.

Pyrko, I., Dörfler, V. and Eden, C. (2017). Thinking together: what makes communities of practice work? *Human Relations*, **70**(4). SAGE Publications Sage UK: London, England: 389–409.

Ribes, D. (2014). Ethnography of scaling, or, how to fit a national research infrastructure in the room. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. pp. 158–70.

Ribes, D. and Baker, K. (2007). Modes of social science engagement in community infrastructure design. *Communities and Technologies 2007*. Springer, pp. 107–30.

Star, S. L. (1999). The Ethnography of Infrastructure. *American Behavioral Scientist*, **43**(3): 377–91

Wenger, E. (1999). *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press.

Distant Reading of the German Coalition Deal

Zylla, Michael
michael.zylla@stud.uni-goettingen.de
University of Göttingen, Germany

Haider, Thomas Nikolaus
thomas.haider@uni-goettingen.de
University of Göttingen, Germany; Max Planck Institute for Empirical Aesthetics, Frankfurt

Motivation

In postwar Germany, the federal government is usually formed by several political parties (Schmidt, 2007, p. 97). Over the past 16 years, these government coalitions were led by the Christian Democratic parliamentary group (CDU/CSU), most recently in cooperation with the Social Democratic Party (SPD), which, following the federal election in 2021 was unwilling to negotiate with their former partner, calling for new alliances to achieve a majority in parliament. Finally, the leaders of the Free Democratic Party (FDP), the Greens and SPD, despite mixed support from the party bases, signed a coalition agreement. Some journalists even regarded the FDP, which gained access to two key ministries, the secret winner of the negotiations (Fürstenau, 2021), also because the Greens did not see some of their desired climate change policies implemented (Lauter, 2021),

In this research, we are interested in how the coalition agreement was assembled regarding the individual party contributions. To that end, we utilize methods from Natural Language Processing, which have seen widespread adoption in political science (Wilkerson and Casas, 2017; Merz et al., 2016; Rauh, 2015; Slapin and Proksch, 2008). Specifically, we carry out a text classification task with transformer models, based on paragraphs from the party manifestos, and use the resulting model to characterize the coalition deal.

Data

Our data consist of the election manifestos from 2021 of the six parliamentary parties, namely Alternative for Germany (AfD), FDP, Greens (Grüne), Left (Linke), SPD, and CDU/CSU (Union), and also the final coalition deal. We converted the original PDFs to plaintext, removed the tables of contents, cleaned the texts from formatting artefacts, and segmented the documents into individual paragraphs. As seen in Table 1 and Figure 1, both document and paragraph length vary widely. The manifestos of AfD and SPD in particular are fairly short, when compared to the Greens and the Left.

Party	Document Length	Number of Paragraphs
AfD	28,171	674
FDP	37,710	493
Grüne	74,757	369
Linke	76,308	1,473
SPD	26,553	386
Union	46,669	1,228

Table 1: Size of German party manifestos.

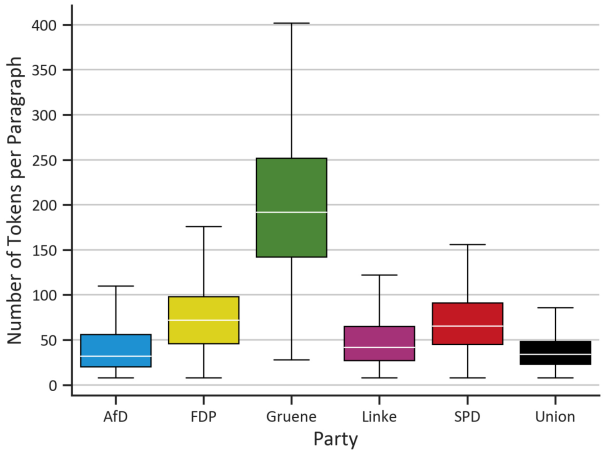


Figure 1: Paragraph length in manifestos

Experiments

To investigate the composition of the coalition deal, we trained German BERT (Devlin et al., 2019) models on a text classification task with the paragraphs of the party manifestos. We test two models: (1) A classification with all six parties, and (2) a classification with the three coalition partners. We examine the difficulty to classify individual parties, how they are misclassified for each other with confusion matrices, and how confident the classifier is w.r.t.

certain paragraphs with a softmax layer (e.g., a paragraph could be assigned 50% SPD, 30% Greens and 20% FDP). Finally, we apply the three-party model to the coalition deal to analyze its composition.

Results

We find that the six-class model has markedly more problems recognizing SPD and Union (see Table 2). Furthermore, SPD paragraphs are often misclassified as Union, while the inverse is not as frequent (see Figure 2). This might be because both are Germany's largest catch-all-parties, with the other parties having a more distinct vocabulary. Altogether, much of the models' confusion falls in line with overlapping political positions, e.g., paragraphs from FDP and AfD are misclassified as Union, while the Greens are never mistaken for the former two. On the other hand, Linke and AfD may be mistaken for each other due to their oppositional language. Thus, it is an open question to what extent the model makes decisions based on policy or language.

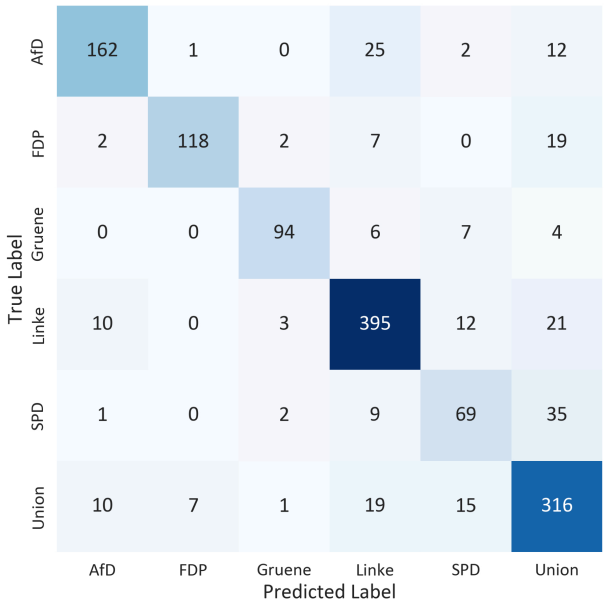


Figure 2: Confusion Matrix of Second model

	precision	recall	f1-score	support
AfD	0.8757	0.8020	0.8372	202
FDP	0.9365	0.7973	0.8613	148
Gruene	0.9135	0.8559	0.8837	111
Linke	0.8565	0.8934	0.8746	441
SPD	0.6699	0.5948	0.6301	116
Union	0.7770	0.8614	0.8170	368
accuracy			0.8333	1386
macro avg	0.8382	0.8008	0.8173	1386
weighted avg	0.8357	0.8333	0.8327	1386

Table 2: Evaluation of First model

The three-class model achieves better classification results (see Table 3), which is not surprising, since the task is easier with less parties to choose from. However, SPD paragraphs are still harder to predict.

	precision	recall	f1-score	support
FDP	0.9026	0.9456	0.9236	147
Gruene	0.9189	0.9189	0.9189	111
SPD	0.8532	0.8017	0.8267	116
accuracy			0.8930	374
macro avg	0.8916	0.8887	0.8897	374
weighted avg	0.8921	0.8930	0.8921	374

Table 3: Evaluation of Second model

As can be seen in the confusion matrix (Figure 3), the model can reliably distinguish FDP and Greens, but both

are harder to distinguish from SPD. This, again, might be explained by the SPD's claim to being a catch-all party.

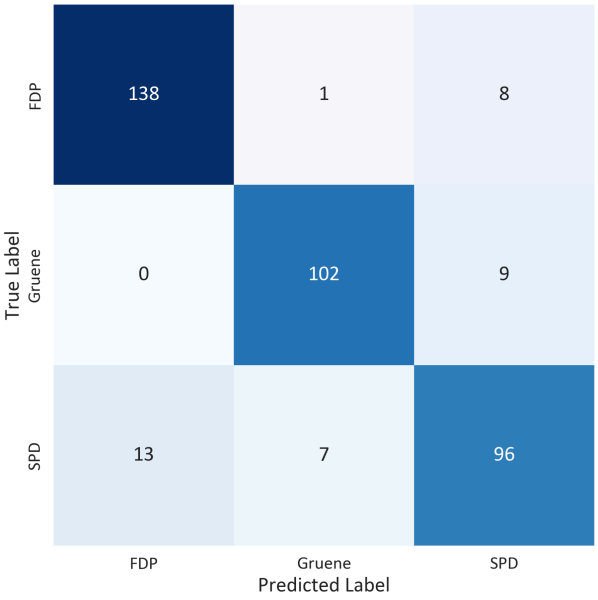


Figure 3: Confusion Matrix of First model

Lastly, we apply the three-class model to the coalition agreement. Figure 4 shows that the model attributes almost 80% of all paragraphs to the SPD. This could be interpreted such that the SPD emerged as the winner of the negotiations. However, this result also reflects the low recall of SPD (Table 3), where the model wrongly tends to classify a paragraph as SPD. Yet, close reading showed that the model's certainty (softmax) was quite high (>99%) for numerous paragraphs, even if they could sensibly be attributed to multiple parties (e.g., in the case of minimum wage and unemployment benefits). Paragraphs with a low certainty were fairly infrequent, and mostly composed of language that is not policy critical.

Finally, keeping in mind the parties share of votes, we would have expected the Greens' proportion to be larger than that of the FDP. Instead, the latter slightly outnumbers the former. More research is needed to disseminate the roles of the smaller parties in the coalition agreement (e.g., binary classification), and also regarding an explanation of the model's decisions.

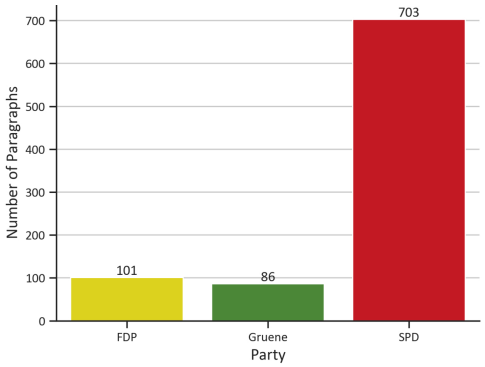


Figure 4: Classification results for paragraphs in coalition agreement (second model)

Bibliography

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (Proceedings of NAACL-HLT): 4171–86.

Fürstenau, M. (2021). How the smallest party in Germany's new coalition came out on top in governing deal *DW.COM* https://www.dw.com/en/how-the-smallest-party-in-germanys-new-coalition-came-out-on-top-in-governing-deal/a-59972370 (accessed 20 April 2022).

Lauter, R. (2021). Ampel-Koalition: Grüne stimmen für Koalitionsvertrag mit SPD und FDP. *Die Zeit*. Hamburg, sec. Politik https://www.zeit.de/politik/deutschland/2021-12/gruene-stimmen-fuer-koalitionsvertrag-mit-spd-und-fdp (accessed 20 April 2022).

Merz, N., Regel, S. and Lewandowski, J. (2016). The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis. *Research & Politics*, **3**(2): 205316801664334 doi: 10.1177/2053168016643346.

Rauh, C. (2015). Communicating supranational governance? The salience of EU affairs in the German Bundestag, 1991–2013. *European Union Politics*, **16**(1): 116–38 doi: 10.1177/1465116514551806.

Schmidt, M. G. (2007). *Das politische System Deutschlands: Institutionen, Willensbildung und Politikfelder*. Orig.-Ausg. (Beck'sche Reihe 1721). München: Beck.

Slapin, J. B. and Proksch, S.-O. (2008). A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science*, **52**(3): 705–22 doi: https://doi.org/10.1111/j.1540-5907.2008.00338.x.

Wilkerson, J. and Casas, A. (2017). Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges. *Annual Review of Political Science*, **20**(1): 529–44 doi: 10.1146/annurev-polisci-052615-025542.

Author Index

Abraham, Vijoy 614
Acosta, Ines 340
Adamou, Alessandro 329
Agata, Mari 415
Agata, Teru 415
Al Moubayed, Noura 90
Alassi, Sepideh 416
Albury, Lauren 700
Algee-Hewitt, Mark Andrew 109
Altin, Ersin 615
Ambarani, Tejas 87
Anderson, Deborah {Debbie} 616
Andresen, Melanie 57
Andrews, Tara Lee 111, 618
Anđelović, Aleksandar 618
Antoniak, Maria 113
Antonini, Alessio 417
Aono, Michihiko 619
Argamon, Shlomo 339
Arnold, Frederik 418
Arnold, Matthias 233
Arul, Kumaran 614
Aubry, Mathieu 45
Auracher, Jan 421
Autiero, Serena 422
Babeu, Alison 186
Bahier-Porte, Christelle 65
Baillot, Anne 115
Bainbridge, David 620
Bajena, Igor Piotr 117
Barabucci, Gioele 680
Barbosa, Denilson 342
Barbot, Laure 68
Barget, Monika 564
Barker, Elton 76
Barnard, Sara 317
Baroncini, Sofia 120
Barré, Jean 471
Barth, George 614
Baude, Olivier 513
Baumard, Nicolas 285
Becker, Devin 73, 599
Beelen, Kaspar 122, 128
Bell, Peter 90, 284, 409
Bench, Harmony 621
Benda, Yuh-Fen 526
Ben-Gigi, Nati 424
Benito-Santos, Alejandro 124
Bergel, Giles Edward (Giles) 27, 65
Bernasconi, Valentine 425
Bernhard, Delphine 557
Berrizbeitia, Francisco 152
Beshero-Bondar, Elisa Eileen 28, 126
Biber, Hanno 623
Birkenes, Magnus Breder 427
Birkholz, Julie 624
Bisceglia, Marta R. 539
Black, Carolyn 496
Blankenship, Avery 98
Bläß, Sandra 637
Bleeker, Elli 128
Blessing, Andre 591
Blicher Christensen, Mathilde 238
Blombach, Andreas 130
Bohmann, Axel 83
Bologna, Federica 134
Bonch-Osmolovskaya, Anastasia 136
Bonora, Paolo 575
Boot, Peter 94
Bordalejo, Barbara 30, 324, 428, 701
Bordonaba-Plou, David 140
Borgia, Mia 127
Börner (Boerner), Ingo 32, 624
Bosse, Arno 430
Bourgeois, Nicolas 542
Brandes, Phillip 70
Brata Roy, Samya 101
Brenner, Simon 431
Broadwell, Peter 141, 614
Broeder, Daan 68
Brottrager, Judith 432
Brown-DeVost, Bronson 690
Brown, Susan 342, 496
Bruschke, Jonas 527
Buchanan, George 300
Burghardt, Manuel 144, 215, 654
Burkert, Mattie 73
Burrows, Toby 627
Busch, Anna 628
Bustamante, Alexandre 152
Byszuk, Joanna 239
Calvo Tello, José 147
Camlot, Jason 151
Campagnolo, Alberto 433
Camps, Jean-Baptiste 155, 160, 285
Can, Yekta Said 165
Cantera, Alberto 665
Capucao. Jr., Reynaldo Caasi 435
Carver, Jeffrey 633
Casties, Robert 633, 634
Cetinic, Eva 90
Chagué, Alix 437
Chaillou, Christelle 155
Chambers, Sally 128, 624

Chan, Andrew Marcus 601
Chan, Jacqueline 127
Chao, Jo-Yu 168
Chardon, Laurette 471
Charvat, Vera 624
Chateau-Dutier, Emmanuel 689
Chatzipanagiotou, Marita 519
Chen, Annie T. 697
Chen, Jing 385, 516
Chen, Song 569
Chen, Wu Wei 171
Chen, You-Jun 172, 696
Chen, Yuqi 440
Chiang, Yao-Yi 444
Chiarcos, Christian 176, 179, 445
Chiffoleau, Floriane 437
Christie, Jennifer 43
Christlein, Vincent 409
Chuang, Tyng-Ruey 104
Cinková, Silvie 624
Ciotti, Fabio 181
Clanuwat, Tarin vii
Clavert, Frédéric 35, 183
Clérice, Thibault 447
Colditz, Iris 665
Cole, Camille Lyans 697
Coll Ardanuy, Mariona 37
Concordia, Cesare 512
Connell, Sarah 79, 98
Conroy, Melanie 449
Costiner, Lisandra 536
Crane, Gregory 185
Crespi, Serena Carlamaria 631
Crompton, Constance 451, 581
Croxall, Brian 188
Cruz, Frances Antoinette 189
Cuéllar, Álvaro 348
Cummings, James 191, 496
Dabrowska, Monika 452
Dahiya, Lavanya 355
Dahiya, Vasundhra 355
van Dalen-Oskam, Karina 625, 636
Damasah, Elliot 51
Damerow, Julia 633, 634
Dangoisse, Pascale 451
Daquino, Marilena 120
Das, Rajarshi 468
Dayter, Daria 83
Daza, Angel 39
De Deyne, Simon 582
Deierl, Marin 111
Dejaeghere, Tess 625
Dekel, Yael 635
Denis, Loïc 65
Dennerlein, Katrin 70, 193
Díaz, Aitor 351
Dickey, Tierany 700
Diehr, Franziska 258
Dietz, Katharina 552
Dodd, Maya 79
Dombrowski, Quinn 101, 532
Donlan, Lisa 83
Dooley, Chase 584
Doran, Michelle 124
Dörk, Marian 76
Downie, J. Stephen 43, 620, 658, 674
Drach, Mortimer 446
Drobac, Senka 254
Drucker, Johanna 76
Drury, Lindsey 258
Duan, Yunxin 593
Dubnicek, Ryan (Ryan C.) 43, 302, 658, 674
Dudar, Julia 625
Dumoulin, Pierre Gabriel 291
Ďurčo, Matej 68, 406, 625
Dutta, Abhishek 27, 65
Dworak, Daniel 117
Earhart, Amy 82, 197
Ebel, Carla 111, 618
Eder, Maciej 199, 453, 625, 636
Edmond, Jennifer 124, 625
Egloff, Mattia 329
Eide, Øyvind 665
Elisar, Ori 36
Elkobi, Jonathan 455
Elo, Kimmo 254
Elswit, Kate 622
Elwert, Frederik 422
Emanuel, Chagai 665
Emery, Doug 434
Emonet, Rémi 65
Enevoldsen, Kenneth 281
van Erp, Marieke 358
Esprit, Schuyler Kirshten 295
Essam, Bacem 457
Evert, Stephanie 130, 202
Fafalios, Pavlos 205
Fan, I-Chun 696
Fäth, Christian 176
Fauchié, Antoine 54, 458
Feng, Minxuan 605
Fialho, Olivia 94
Fickers, Andreas 184
Fiechter, Benjamin 419
Fields, Sam E. 697
Fileva, Evgeniia 625
Fischer, Frank 32, 68, 406, 486, 625
Flüh, Marie 368, 459, 637

Focht, Josef 274
Fokkens, Antske 40
Fonteyn, Lauren 461
Fossati, Alessandro 94
Fournel, Thierry 65
Freedman, Richard 563
Fricke, Heike 274
Frizzera, Luciano 342
Frühwirth, Timo 431
Fu, Yaming 462
Fuhry, David 274
Fujimaki, Satoshi 648, 671
Fukumoto, Takaki 661
Fushimi, Takeshi 506
Galina Russell, Isabel 464
Galleron, Ioana 471
Gandham, Sukesh 152
Gangemi, Aldo 358
Gao, Andrew 633, 634
Garcia, Noa 45
Garrobo-Peral, Manuel 550
Gatti, Tommy 466
Gautrais, Thomas 65
Gawley, James 207
Gebhard, Henning 552
Gelumbeckaite, Jolanta 445
Ghosh, Arjun 209
Ghosh, Sharanya 468
Gil, Alex 73
Gil Fuentes, Alexander 116
Giovannini, Luca 639
Girard, Paul 264
Gittel, Benjamin 476
Gius, Evelyn 419, 473, 588
Gladstone, Clovis 382
Glass, Grant 210
Glover, Kaiama L 116
González-Blanco, Elena 352
Gooding, Paul 321
Goodwin, Paul 488
Gordin, Shai 212
Goto, Makoto 499, 610
Graham, Maryemma 658
Grallert, Till 346
Grandjean, Martin 641
Gray, Edward 68, 512
Grillo, Remo 575
Grisot, Giulia 239, 469
Grunspan, Claude 471
Guénette, Marie-France 101
Guerard, Elisabeth 36
Guhr, Svenja Simone 472
Gunn, Stan 660
Gutiérrez de la Torre, Silvia E. 215
Habrard, Amaury 65
Hadden, Richard William James (Richard) 40, 474
Van Hage, Willem Robert 94
Hagen, Thora 270
Haggin, Patience 642
Haider, Thomas Nikolaus 476, 703
Hammond, Adam 218, 220
Han, Zhitong 602
Hankinson, Andrew 560
Hannah, Matthew Nathan 79
Hannesschläger, Vanessa 224
Hansen, Lasse 281
Hara, Shoichiro 104
Hara, Shoko 645
Hartinger, Teresa 478
Hashimoto, Emi 648
Hashimoto, Takako 388
Hashimoto, Yuta 479
Haverinen, Jonas 489
Hayden, Gabriele 599
Hayward, Nicholas John 483
He, Daqing 225
Heaton, Raina 301
Heibi, Ivan 484
Heiden, Serge 625
Hein, Pascal 591
Helling, Patrick 227, 230, 680
Henke, Konstantin 233
Hermsen, Lisa 236
Herold, Nastasia 238
Herrmann, J. Berenike (Berenike) 94, 239, 469
Hervieux, Natalie 342
Heßbrüggen-Walter, Stefan 486
Hibino Lory, Harumi 430
Hidalgo Urbaneja, Maribel 488
Hiippala, Tuomo 489, 576
Hilbing, Genna 620
Hiltmann, Torsten 491
Hinzmann, Maria 552
Hirano, Michiru 691
Hirst, Graeme 221
Hodošček, Bor 647
Holmes, Martin 241
Holmes-Wong, Deborah 444
Hoppe, Stephan 527
Hosseini, Kasra 37
Hotho, Andreas 684
Hou, Yumeng 329
Houston, Natalie 492
Hsiang, Jieh viii
Hsieh, Hsin-Yi 168, 172
Hu, Yuerong 620
Huang, Shi-Yun 168
Huber, Alexander 243

Huber, Emma 243
Huijnen, Pim 246, 249
Humbel, Marco 252
HUNG, JEN JOU 494
Hung, Ying-Fa 104
Hyvönen, Eero 40, 254
Iashchenko, Anatoly Vladimirovich 256
Iezzi, Adriana 539
Ikkala, Esko 254
Illmer, Viktor J. 257
Ilovan, Mihaela 342, 496
Imamura, Satoshi 404
Impett, Leonardo 45, 89
Inoue, Satoshi 663
Inoue, Sayaka 479
Ionov, Maxim 176
Ishibashi, Keiichi 672
Ishii, Tatsuya 671
Ishikawa, Kazuki 661
Isogai, Kana 269
Isuster, Marcela Y. 495, 496
Ivan, Loredana 48
Ivanov, Lubomir 262
Iwasaki, Junya 661
Jablonski, Pawel 690
Jacke, Janina 70
Jacomy, Mathieu 641
Jakacki, Diane Katherine 496
James, Stuart 45
Jang, Jr-Jie 104
Jänicke, Stefan 238, 310
Jannidis, Fotis 130, 270, 279
Jäschke, Robert 419
Jiang, Ming 620
Jimenez-Mavillard, Antonio 622
Johnsen, Lars G. 427
Johnson, David F. 264
Jolivet, Vincent 447, 497
Jorschick, Annett 94
Joseph, Ben 152
Joyce, Terry 583
Jreis-Navarro, Laila M. 140
Jügel, Thomas 665
Jung, Kerstin 230
Juola, Patrick 267
Kabadayi, M. Erdem 165
Kacsuk, Zoltan 354
Kahmann, Christian 667
Kamata, Ryo 404
Kameda, Akihiro 499
Kanagawa, Nadia 55
Karadkar, Unmil 48
Kåsen, Andre 427
Kato, Takahiro 648
Katzoff, Binyamin 424
Kawase, Akihiro 268
Kawazu, Kosei 671
Kenderdine, Sarah 329
Kesäniemi, Joonas 254
Kestemont, Mike 189
Ketchley, Sarah 500
Ketzan, Erik 270
de Keulenaar, Emillie Victoria 501
Keydar, Renana 305
Khulusi, Richard 274
Kim, Boyoung 479
Kim, Eric 101
Kim, Hoyeol 502
Kim, Jina 444
Kisjes, Ivan 501, 503
Kitamoto, Asanobu 276, 578
Kizer, Ianna 700
Kizhner, Inna 315
Klee, Anne 552
Kleemola, Mari 68
Kleindienst, Nina 649
Kleymann, Rabea 215, 505
Klic, Lukas 575
Kluvanec, Dan 90
Kobayashi, Michikazu 692
Kobayashi, Ryota 388, 695
Koho, Mikko 40, 254
Kokaze, Naoki 479, 506
Kölligan, Daniel 530
König, Alexander 68
Konle, Leonard 130, 278
Konstanciak, Johanna 552
Koolen, Marijn 94, 128
Kosti, Ronak 90, 284
Krautter, Benjamin 57
Křen, Michal 625
Kriebernegg, Ulla 48
Kristensen-McLachlan, Ross Deans 281
Kriukov, Artem 136
Kroll, Simon 508
Kröncke, Merten 278
Kroon, Ariel 87
Kudeki, Deren 43
Kulkarni, Kavita 509
Kulyabin, Mikhail 284
Kumakura, Wakako 510
Kunda, Bartłomiej 625
Kurita, Kazuhiro 695
Kuroczyński, Piotr 117
Kuroshima, Satoru 663
Kurzynski, Maciej 597
Kuzman-Slogar, Koraljka 224
LaCelle-Peterson, Nathaniel 76

La Mela, Matti 254
Land, Kaylin Catherine (Kaylin) 50, 342
Langlais, Pierre-Carl 285
Lapin, Hayim 690
Larrousse, Nicolas 512, 513
Lassen, Ida Marie S. 281
Lastilla, Lorenzo 522
Law, Shun-Man 601
Lawrence, Dan 90
Lawrence, Jon 123
Layne-Worthey, Glen 43, 620, 674
Leal, Rafael 254
Lee, Benjamin Charles Germain 289
Lee, Cheng-Jen 104
Leelasorn, Angel 51
Leemans, Inger 587
Lefranc, Lith 90
Lehmann, Jörg 348
Lemke, Marc 368, 459
Lensink, Saskia 636
Leone, Anna 90
Lescouet, Emmanuelle 291, 514, 653
Leskinen, Petri 254
Leuckert, Sven 84
Lewis, David 560
Li, Bin 605
Li, Mengqi 516
Li, Yi 292
Li, Yongning 293
Li, Zekun 444
Liao, Hsiung-Ming 696
Licastro, Amanda Marie 82, 295
Liebe, Lauren 52, 73, 296
Liebl, Bernhard 654
Liimatta, Aatu 84
Lin, Hong-Ren 298
Lin, Wensi 516
Lipski, Candice 451
Liu, Chao-Lin 298
Liu, Guanwei 656
Liu, Rui 300
Liu, Wei 463
Liu, Wei-Zhi 298
Liu, Yidan 536
Liu, Yi-Fan 298
Losh, Elizabeth 82, 518
Lu, Kun 301
Lu, Xuehui 605
Ludwig, Jess 500
Luhmann, Jan 144
Luu, Thi Kim Hanh 302
Luyk, Sean 87
Ma, Rongqian 225, 549
McConnell, Kyla 84
Maccori Kozma, Gustavo 83
MacDonald, Andrew 50, 342
McDonough, Katherine 123, 444
Machotka, Ewa 519
McKay, Dana 300
Madhu, Prathmesh 409
Maehara, Noriko 672
Maeir, Noam 305
Magni, Isabella 302, 658, 674
Mahler, Hanna 84
Mahony, Simon 463
Maier, Andreas 409
Maiwald, Ferdinand 527
Mallen, Enrique 525
Mandell, Laura 52, 296
Manjavacas Arevalo, Enrique 307, 461
Mapp, Rennie 660
Mariani, Fabio 520
Marinšek, Urša 478
Mariotti, Viola 155
Marshall, Sophie 70
Marsocci, Valerio 522
Martínez Nieto, Roxana Beatriz 308
Martus, Steffen 419
Massanari, Adrienne 84
Matsuda, Tomohiro 695
Matusiak, Krystyna K 389
Mayer, Sandra 431
Meier-Vieracker, Simon 486
Meinecke, Christofer 310
Mélanie, Frédérique 471
Melga-Estrada, Liliana 128
Mellet, Margot Lise 54
Meneses, Luis 524, 525
Menini, Stefano 587
Menninghaus, Winfried 421
Merenda, Martina 539
Mertgens, Andreas 387
Messerli, Thomas C. 83, 94
Messina, Cara Marta 98
Michney, Todd 340
Middle, Sarah 313
Miller, Tracy 526
Milling, Carsten 32
Mimno, David 113, 134
Minster, Sara 315
Miya, Chelsea 87
Miyagawa, So 671
Miyao, Yusuke 583
Mohammad, Saif M. 220
Molitor, Paul 302
Mondaca, Francisco 665
Moore, Shawn 73
Mori, Shinsuke 610

Morin, Olivier 285
Morini, Francesca 76
Morrison, Zachary 87
Morrissey, Robert 344
Moscatelli, Cristiano 422
Mousavi, Emad 317
Mrugalski, Michał 625
Mu, Yu-Chia Monica 104
Muehlberger, Guenter 322
Muenster, Sander 527
Münster, Sander 117
Murai, Hajime 661
Murphy, Ciara 625
Murphy, Orla 529, 535
Nagasaki, Kiyonori 28, 479, 645, 668, 698
Naguib, Marco 471
Nakamura, Satoru 479, 656, 663, 668, 687
Nakamura, Shougo 661
Nakamura, Yusuke 506
Nakanishi, Yasuhito 672
Nanditha, Narayanamoorthy 318, 320
Nanni, Federico 37
Natsume, Muneyuki 404
Nemoto, Sakura 661
Neuefeind, Claes 530, 665
Neugarten, Julia 94, 428
Neugebauer, Tomasz 151
Newton, Greg 241
Niccolucci, Franco 580
Nicosia, Marissa 73
Niebling, Florian 527
Niekler, Andreas 144, 215, 667
Nielbo, Kristoffer L. 281
Nishi, Hironori 449
Nishimura, Taichi 610
Nockels, Joseph Hiliary 321
Nomura, Nichole Misako 532
van Noord, Nanne 45
Nurmikko-Fuller, Terhi 466
Odebrecht, Carolin 625
O'Donnell, Daniel 30, 323, 701
O'Driscoll, Michael 87
Offert, Fabian 89
Ogawa, Jun 668
Ohba, Arisa 661
Öhman, Emily Sofi 533
Ohmukai, Ikki 645, 668, 692
Ohta, Shoki 661
Oka, Toshio 671
Okada, Kazuhiro 670
Okada, Takashi 648, 671
Olsen, Mark 344
Ope-Davies (Opeibi), Tunde 79
Organisciak, Peter 389
Orr, Raymond 301
Osawa, Tomejiro 648, 672
O'Sullivan, James 535
Otis, Jessica 73
Page, Kevin 560
Page-Perron, Emilie 179
Pagel, Janis 57, 326
Pai, Pi-Ling 696
Pala, Giovanni Maria (Giovanni) 536, 538
Palladino, Chiara 55
Papadopoulos, Costas 364
Park, Juyong 596
Parulian, Nikolaus Nova 674
Pasqual, Valentina 120, 539
Pattee, Aaron 527
Pavlopoulos, John 519
Peaker, Alicia 116
Pellen, Marie 513
Pellet, Aurélien 542
Pérez, Álvaro 352
Perkins, Patrick 434
Peroni, Silvio 485
Pfeffer, Magnus 354
Pfeiffer, Mirjam 36
Pianzola, Federico 94
Picca, Davide 329, 597
Pichler, Axel 57
Pickering, Paul 466
Pidd, Michael 535
Pielström, Steffen 70, 130, 230
Pierazzo, Elena 332
Pilla, Julien 447
Pinche, Ariane 160, 676
Piontkowitz, Vera 274
Plancq, Clément 471
Plets, Gertjan 249
Pöckelmann, Marcus 302
Podriadchikova, Maria 136
Poggel, Lisa 258
Poibeau, Thierry 471
Ponce de la Vega, Lidia 101
Pons, Jessie 422
Porter, Dot 433
Porter, J.D. 333
Povroznik, Nadezhda 79
Prajda, Katalin 618
Priani Saisó, Ernesto 464
Proisl, Thomas 130
Puren, Marie Anna 541
Quinn, William Reed 98
Raciti, Marco 625
Radisch, Erik 678
Ransom, Lynn 434
Rau, Felix 679

Raynor, Cecily 101
Read, Lewis 618
Rebora, Simone 94, 469, 544
Reeder, Jake 335
Rees, Gethin 337
Reinöhl, Uta 530
Reiter, Nils 473
Rezania, Kianoosh 665
Rhodes, Josh 123
Riaño Rufilanchas, Daniel 547, 547
Richards, Nina 618
Riddell, Allen 400
Rinderlin, Jonas 76
Ringler, Hannah 339
del Rio Riande, Gimena 391
Risha, Zak 549
Rittenhouse, Brad 79, 340
Ritter, Jörg 302
Rockwell, Geoffrey Martin (Geoffrey) ..... x, 50, 317, 342
Rod, Alisa B. 496
Rodrigues, Elizabeth Sarah 681
Roe, Glenn 344
Roeder, Torsten 116, 628
Rojo-Mejuto, Natalia 550
Romach, Avital 212
Romanov, Maxim 346
Romary, Laurent 437
Ros, Salvador 352, 625
de la Rosa, Javier 147, 348, 351
Rosenthaler, Lukas 416
Rosol, Christoph 76
van Rossum, Lisanne 625
Roth, Martin 354
Röttgermann, Julia 552
Roux, Dominique 513
Roy, Dibyadyuti 295, 355, 556
Rózsa, Márton 618
Ruiz Fabo, Pablo 557
Saccomano, Mark 560
Sack, Graham Alexander 357
Saelid, Daniel P. 697
Sahle, Patrick 530
Saito, Yuuri 661
Sakamoto, Shouji 682
Salgaro, Massimo 544
Sanders, Ashley 561
Sanfilippo, Emilio M. 563
Santa Maria, Teresa 452
Sartini, Bruno 358
Saviotti, Federico 155
Sawchuk, Kim 48
Scharinger, Mathias 421
Scheithauer, Hugo 437
Schler, Jonathan 424
Schlesinger, Claus-Michael 591
Schlögl, Matthias 474
Schlör, Daniel 684
Schmidt, Thomas 193, 649
Schneider, Felix 70
Schneider, Philipp 491
Schneider, Stefanie 361
Schöch, Christof 552, 625
Scholger, Martina 28
Scholger, Walter 116, 224
Schreibman, Susan 364, 564
Schroeter, Julian 686
Schumacher, Mareike Katharina 365, 368
Schwartz, Michelle 451
Šeļa, Artjoms 199, 625
Selisker, Scott 566
Seminck, Olga 471
Seung-Bin, Yim 406
Shamsian, Farnoosh 55, 186
Shang, Wenyi 567, 569
Sharma, Srishti 94
Shaw, Robert L. J. 370
Shibata, Elisabete 560
Shibutani, Ayako 687
Shigehara, Toru 479
Shiki, Yoko 672
Shimoda, Masahiro 698
Shinozaki, Yuji 660
Shiratori, Takayuki 661
Sibeko, Johannes 571
Simard, Benoît 264
Simon, Rainer 444
Sinatra, Michael 689
Sinikallio, Laura 254
Sluyter-Gäthje, Henny 32, 372, 573
Smits, Thomas 90
Smolarski, René 117
Spinaci, Gianmarco 575
Srinivasan, Venkat 556
Stapel, Rombert 375
Stauder, Andy 322
Steffes, Moritz 552
Stern, Simon 218
Stokes, Peter Anthony 58, 378
Stökl Ben Ezra (Stoekl Ben Ezra), Daniel 58, 690
Streiter, Oliver 104
Stringfield, Ravynn K. 82
Sturgeon, Donald 61, 380
Stutzmann, Dominique 160
Suda, Makiko 663
Suda, Towa 695
Sunada, Kyosuke 672
Sunaga, Emiko 510
Sunuwar, Dev Kumar 616

Sutton Koeser, Rebecca 633, 634
Suviranta, Rosa 576
SUZUKI, Chikahiko 578
Suzuki, Kazunori 691
Swift, Ben 466
Tadjo Takianpi, Yves 437
Tagliaferri, Lisa 63, 63, 580
Taipale, Sakari 48
Takagi, Soichiro 645
Takahashi, Mai 692
Takata, Yuichi 580
Takedomi, Yuka 388, 695
Tamper, Minna 254
Tangherlini, Timothy R. 141
Taniguchi, Chikamitsu 648
Taparia, Kanishtha 556
Tarpley, Bryan 52
Tasovac, Toma 625
Tauber, James 186
Tayler, Felicity 581
Tchoc, Bennett Kuwan 50
Telegina, Maria 582
Terras, Melissa 292, 321
Terriel, Lucas 437
Thalken, Rosamond 134
Tharsen, Jeffrey 382, 633, 634
Therón, Roberto 124
Thiel, Carsten 68
Thomas, Drew 65
Thomas, Grace 584
Thorat, Dhanashree 585
Thornhill, Kate 599
Tomasi, Francesca 120, 539
Tomonari, Yūki 648
Tong, Wei 593
Tonra, Justin 625
Torres, Sergio 497
Tosin, Rafaela 84
Tóth-Czifra, Erzsébet 625
Toyosawa, Shuuhei 661
Trettien, Whitney 586
Trilcke, Peer 32, 372, 625
Tsai, Jung-Yi 696
Tsai, Richard Tzong-Han 168, 172, 696
Tse, Devi 601
Tsui, Lik Hang 385
Tullett, William 587
Tuominen, Jouni 254
Türkoglu, Enes 386
Uglanova, Inna 588
Ulrich, Mona 591
Underwood, Ted 567
Uno, Takeaki 388, 695
Utescher, Ronja 527
Uzor, Tia-Monique 622
VandenBosch, Adrienne 389
Varner, Stewart 333
Vauth, Michael 419
Velios, Athanasios 488
Ventresque, Vincent 65
Verdini, Paolo 317
Verheul, Jaap 249
Verhoeven, Deb 627
Vernet, Marguerite 160
Vernus, Pierre 542
Vetter, Alyssa 301
Vial-Bonacci, Fabienne 65
Vidal-Gorène, Chahan 160
Viehhauser, Gabriel 591
Viglianti, Raffaele 127, 391
Viola, Lorella 393
Vishnubhotla, Krishnapriya 220
Vitale, Valeria 37, 444
Vitali, Fabio 120
Vitali-Rosati, Marcello 54, 515, 653, 689
Viviani, Marco 94
Vogeler, Georg 474
Vogelmann, Valentin 358
Vogl, Malte 633, 634
Vorobieva, Viktoria 136
Wacławik, Paulina 396
Walker, Rebekah 236
Walsh, John A. 43, 658
Walsh, Melanie 113
Walton, Jo Lindsay 116
Wandl-Vogt, Eveline 40
Wang, Changsong 593
Wang, Chia-Hsun Ally 104
Wang, Haining 400
Wang, Jingwei 317
Wang, Jisheng 516
Wang, Jun 593
Wang, Linxu 440, 593
Wang, Shuofei 536
Wang, Sungpil 596
Wang, Yadi 697
Wang, Yifan 698
Wang, Yiwen 597
Wang, Yu-Chun 494
Wang, Yu-Huang 104
Watson, Matthew 90
Waxman, Joshua 397
Wegner, Jacob 186
Werner, Carole 557
Wessels, Bridgette 535
West-White, Clarissa 674, 700
Westerling, Kalle 633
Wevers, Melvin 90

Wicht, Bertil 597
Wieder, Philipp 406
Wikle, Olivia M. (Olivia) 73, 599
Wiles, Simon 614
Wilkens, Matthew 134
Wilkinson, Hazel 65
Williams, Seretha D. (Seretha) 674, 700
Williamson, Evan Peter 73, 599
Wilson, Daniel C.S. 123
Winko, Simone 279
Winterbottom, Tom 90
Wintergruen, Dirk 77
van Wissen, Leon 503
Witt, Andreas 270
Wolff, Christian 193, 649
Wong, Kwong-Cheong 600
Woods, Nathan 30, 323, 701
Worthey, Glen C. 658
Wrisley, David Joseph 310
Xiao, Shuang 529
Xie, Xin 400
Yamada, Taizo 656, 663, 687
Yamamoto, Hilofumi 647, 692
Yanase, Peter 580
Yanbe, Koki 687
Yang, Yuchen 602
Yano, Keiji 404
Yoshida, Takumi 661
Yousef, Tariq 55
Yuan, Yiguo 605
van Zaanen, Menno 571, 606
Zafar, Huma 342
Zaghouani, Wajdi 607
Zandbank, Itay 633, 634
Zarei, Alireza 406
Zarrieß (Zarriess), Sina 473, 527
Zhan, Hanna Yaqing 104
Zhang, Haihui 225
Zhang, Mengyue 516
Zhao, Fudie 608
Zheng, Ruoyun 225
Zhitomirsky-Geffet, Maayan 315, 424
Zhu, Jieyong 610
Zhuge, Jing 526
Zinnen, Mathias 409
Zisserman, Andrew 65
van Zundert, Joris J. (Joris) 128, 428, 636
Zuo, Lala 171, 526
Zylla, Michael 703



東京大学
THE UNIVERSITY OF TOKYO