JADH 2017

Proceedings of the 7th Conference of Japanese Association for Digital Humanities

"Creating Data through Collaboration"

Japanese Association for Digital Humanities

ALLIANCE OF DIGITAL HUMANITIES ORGANIZATIONS

JADH2017

Proceedings of the 7th Conference of
Japanese Association for Digital Humanities

“Creating Data through Collaboration”

http://conf2017.jadh.org/
Doshisha University, September 11-12, 2017

Hosted by:
JADH2017 Organizing Committee
under the auspices of the Japanese Association for Digital Humanities

Co-hosted by:
Department of Culture and Information Science, Doshisha University

Supported by:
Construction of a New Knowledge Base for Buddhist Studies:
Presentation of an Advanced Model for the Next Generation of Humanities Research
(15H05725, Masahiro Shimoda)
International Institute for Digital Humanities

Co-sponsored by:
PSJ SIG Computers and the Humanities
Japan Society for Digital Archive
Japanese Society for Information and Media Studies
Japan Art Documentation Society (JADS)
Japan Association for East Asian Text Processing (JAET)
Japan Association for English Corpus Studies
The Mathematical Linguistic Society of Japan
Japan Society of Information and Knowledge
Alliance of Digital Humanities Organizations

Table of Contents

Session 3:

Poster session:

JADH 2017 Organization

JADH 2017 Organizing Committee:

- Takashi Harada(Doshisha University)
- Sho Sato (Doshisha University)
- Akihiro Kawase (Doshisha University), Chair
- Gen Tsuchiyama (Doshisha University)
- Kiyonori Nagasaki (International Institute for Digital Humanities, Japan)
- Toru Tomabechi (International Institute for Digital Humanities, Japan)

JADH 2017 Program Committee:

- Paul Arthur (Edith Cowan University, Australia)
- James Cummings (Newcastle University, UK)
- J. Stephen Downie (University of Illinois, USA)
- Øyvind Eide (University of Cologne and University of Passau, Germany)
- Neil Fraistat (University of Maryland, USA)
- Makoto Goto (National Museum of Japanese History, Japan)
- Shoichiro Hara (Kyoto University, Japan)
- Jieh Hsiang (National Taiwan University, Taiwan)
- Akihiro Kawase (Doshisha University, Japan)
- Asanobu Kitamoto (National Institute of Informatics, Japan), Chair
- Maciej Eder (Pedagogical University of Kraków, Poland)
- Charles Muller (University of Tokyo, Japan)
- Hajime Murai (Tokyo Institute of Technology, Japan)
- Kiyonori Nagasaki (International Institute for Digital Humanities, Japan)
- John Nerbonne (University of Groningen, Netherlands)
- Geoffrey Rockwell (University of Alberta, Canada)
- Susan Schreibman (National University of Ireland Maynooth, Ireland)
- Masahiro Shimoda (University of Tokyo, Japan)
- Raymond Siemens (University of Victoria, Canada)
- Keiko Suzuki (Ritsumeikan University, Japan)
- Takafumi Suzuki (Toyo University, Japan)
- Tomoji Tabata (Osaka University, Japan)
- Toru Tomabechi (International Institute for Digital Humanities, Japan)
- Kathryn Tomasek (Wheaton College, USA)
- Christian Wittern (Kyoto University, Japan)
- Taizo Yamada (University of Tokyo, Japan)

Time Table

September 11, Day 1

08:30	Registration
09:00-09:15	Opening
09:15-10:45	Session 1
10:45-11:00	Break
11:00-12:15	Session 2
12:15-13:30	Lunch Break
13:30-14:45	Session 3
14:45-15:00	Break
15:00-16:10	Panel session
16:10-16:25	Break
16:25-17:25	Keynote Lecture
17:30-18:30	Poster Slam and Core Time
19:00-	Reception

September 12, Day 2

09:00-10:30	Session 4
10:30-10:45	Break
10:45-12:15	Session 5
12:15-13:30	Lunch break (Annual General Meeting of JADH)
13:30-14:50	Special thematic session
14:50-15:05	Break
15:05-16:35	Session 6
16:35-17:00	Closing

[Keynote lecture]

Collaboration at scale: emerging infrastructures for digital scholarship

Donald Sturgeon
Harvard University

Modern technological society is possible only as a result of collaborations constantly taking place between countless individuals and groups working on tasks which at first glance may seem independent from one another yet are ultimately connected through complex interdependencies. Just as technological progress is not merely a story of ever more sophisticated technologies, but also of the evolution of increasingly efficient structures facilitating their development, so too scholarship moves forward not just by the creation of ever more nuanced ideas and theories, but also by increasingly powerful means of identifying, exchanging, and building upon these ideas.

The digital medium presents revolutionary opportunities for facilitating such tasks in humanities scholarship. Most obviously, it offers the ability to perform certain types of analyses on scales larger than would ever have been practical without use of computational methods – for example the examination of trends in word usage across millions of books, or visualizations of social interactions of tens of thousands of historical individuals. But it also presents opportunities for vastly more scalable methods of collaboration between individuals and groups working on distinct yet related projects. Simple examples are readily available: computer scientists develop and publish code through open source platforms, companies further adapt it for use in commercial systems, and humanities scholars to apply it to their own research; libraries digitize and share historical works from their collections, which are transcribed by volunteers, searched and read by researchers and cited in scholarly works.

Much of the infrastructure already in use in digital scholarship is infrastructure developed for more general-purpose use – a natural and desirable development given the obvious economies of scale which result from this. However, as the application of digital methods in humanities scholarship becomes increasingly mainstream, as digitized objects of study more numerous, and related digital techniques more specialized, the value of infrastructure designed specifically to support scholarship in particular fields of study becomes increasingly apparent. This paper will examine types of humanities infrastructure projects which are emerging, and the potential they have to facilitate scalable collaboration within and beyond distributed scholarly communities.

[Special thematic session]

The Power of Crowdsourcing and Libraries

**Azusa Tanaka (University of Washington),
Takashi Harada (Doshisha University),
Kiyonori Nagasaki (International Institute for Digital Humanities),
Kosetsu Ikeda (Chiba University),
Atsuyuki Morishima (University of Tsukuba)**

Abstract

Crowdsourcing is recognized as a key technology having a big impact on library and digital humanity domains. It has already been used for many purposes such as transcribing texts, creating metadata, and correcting errors in bibliographic records. This session has one keynote and several reports that address ongoing crowdsourcing projects in/for libraries in the United States and Japan and how crowds can help us solve a variety of problems. The talks also discuss important issues found in those projects, potentials and future directions of crowdsourcing and libraries.

Biography

Azusa Tanaka (University of Washington)
Azusa Tanaka is a Japanese Studies Librarian at University of Washington and handles subject liaison, reference, instruction, collection development, and cataloging. More about her research: https://jisao-washington.academia.edu/AzusaTanaka.

Takashi Harada (Doshisha University)
Takashi Harada is a professor at the Center for License and Qualifications in Doshisha University. Aside from the work in the center, he is responsible for the education in the Library and Information Science master' s course as a professor of the Graduate School of Policy and Management in Doshisha University. His research interests include the design and analysis of integrated library systems and social information systems. He is president of Project Next-L, a library community initiative to develop specifications for the open source next generation integrated library system.

Kiyonori Nagasaki (International Institute for Digital Humanities)
Kiyonori Nagasaki is a 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.

Kosetsu Ikeda (Chiba University)
Kosetsu Ikeda is an Assistant Professor at Chiba University. His current main research interests include crowdsourcing, library and information science, graph data theory, and data engineering.

Atsuyuki Morishima (University of Tsukuba)
Atsuyuki Morishima is a Professor of University of Tsukuba, Japan. His research interests include data-centric human-machine computations, data integration, and Web data management. Currently, he is the chair of SIG-DBS of the Information Processing Society of Japan, an editor-in-chief of IPSJ transactions on Database Systems, and an associate editor of the VLDB Journal.

[Panel session]

Transformative Data: Data-Driven Reconfigurations of Interdisciplinary Work in the Digital Humanities

Kristin Allukian (University of South Florida),
Mauro Carassai (California State University, Northridge),
Laura Hale (ParaSport News)

Abstract

Spanning three particular historical moments over the last century and foregrounding different DH approaches to women's visual history, literary theory, and disability sports studies, our interdisciplinary panel addresses the relationship between data collection and gender, subjectivity, and ableism in our DH research, teaching, and across the ever-shifting meaning of "data" within the current political and academic climate. Our papers individually and collectively address the conference theme of crossing borders in digital humanities work by using data analysis as a practice that connects different branches of knowledge. Kristin Allukian's project, which involves the collaboration of undergraduate students, graduate students, and faculty, can be seen as interdisciplinary in nature, crossing various academic fields including literature, women's studies, visual history, and politics. Similarly, Mauro Carassai's paper develops on the threshold between linguistic, literature, and anthropology by looking at the depiction of the human, machinic, and cybernetic subject when the set of unstructured set of data is constituted by an artificially constructed language. Finally, Laura Hale's paper examines how organizations do and do not cross borders when accessing data on disability sports for either commercial or non-profit purposes inviting a general scholarly reflection that involves social sciences, business, and knowledge management. Looking at a series of case studies at different moments in time and across different academic fields, our panels aims to strengthen an understanding of how data-driven inquiries might mobilize past and present dissent, ambiguous borders of inter-cultural linguistic domains, and information accessibility in the research, teaching, and participation of the digital humanities.

Kristin Allukian begins the panel in the early twentieth century, arguing feminist digital humanities practices of archiving pro- and anti-suffrage postcards can engender new visual historical narratives of masculinity, manhood, and fatherhood. Much work has been done on visual representations of women in suffrage postcards; broadly speaking, scholars discuss how the strictures of white heteronormative femininity and the expectations of motherhood were important tropes in suffrage imagery. Much less, however, has been said about how depictions of masculinity and fatherhood operated in suffrage postcards. This project, the Suffrage Postcard Project, utilizes a range of digital tools including Omeka, ImagePlot, Gephi, Tableau Public, and Iconclass to explore how narratives of masculinity and fatherhood contributed to the suffrage debate and fictional literature of the time.

Moving from the visual image to the written page, Mauro Carassai extends the inquiry on representation to the depiction of the human, machinic, and cybernetic subject when the set of unstructured set of data is constituted by an artificially constructed language. Computer-enhanced methodologies of textual analysis can reveal their limits, as well as their advantages, when the encoded semi-structured data undergoing data mining practices are constructed in ways that resist digital discrete differentiation as the creolized Caribbean linguistic set in Nalo Hopkinson's novel Midnight Robber does. Carassai's paper offers a set of different of "distant readings" of the novel operated by means of various software (Voyant, Lexos, NLT) and compares the set's results with textual data parsed by means of stylometric software (Signature, stylo, of Stylene) in order to raise problems in specific theories of marginalized subjectivity that might (and should) influence the future direction of software development for textual literary analysis.

Moving from a forum that is literary is nature to one that is athletic in nature, Laura Hale's project explores the practical issues of creating a large-scale database for disability sport, by

working to create a database of disability sports information. The primary focus is on discussing data accessibility, the decision-making process for the software used, questions around how accessible data should be, and balancing the desire for openness to improve the lives of people with disabilities and improving their sporting performance with trying to monetize these efforts to enable the project to continue. Hale argues that using large datasets, including publicly available datasets that describe general disability characteristics against sporting performance based on classification and gender, can provide greater insight as opposed to the use of smaller scale data analysis that tends to be based around single events.

These three case studies use a wide range of digital tools to remap sites of dissent, identity formation and identity politics, and data accessibility across time and disciplines, providing additional critical networks of gender, cultural, and disability studies —networks that are still in the process of being constructed by means of current digital humanities practices and methodologies.

Individual Papers

1. Kristin Allukian, "The Illustration, the Image, and the Archive: Feminist Digital Humanities Approaches to Caricatures of Masculinity in American Suffrage Postcards, 1900-1920"

In 2017, we have Twitter, Instagram, Snapchat, and Facebook. A hundred years earlier, there were postcards. In the "Golden Age" of postcards (1902-1915), postcards circulated with the same fervor, albeit not speed, of images on popular social media apps today. This project looks back at the early decades of the 1900s in the context of the women's suffrage movement, a movement that was gaining momentum in the same historical moment of the Golden Age of postcards. The Suffrage Postcard Project utilizes a range of digital tools including Omeka, ImagePlot, Gephi, Tableau Public, and Iconclass to explore two overarching questions: 1. how can feminist digital humanities practices engender new visual historical narratives of masculinity, manhood, and fatherhood? And 2. how does this lend an understanding of how such narratives of masculinity and fatherhood contributed to the suffrage debate?

A team of undergraduate students, graduate students, and faculty collaborate to upload and tag postcard images using an approach that has been influenced by Jacqueline Wernimont and Julia Flanders' 2010 article "Feminism in the Age of Digital Archives." In relation to building their digital archive of early modern women's writing, Wernimont and Flanders describe specific challenges by explaining that "[women's] writing often confounds the processes of categorization, explication, and description central to digital text markup…. The work of digitization and encoding also engages us in a reflexive process that forces us to interrogate those genres and any genre-tags that we may use in creating the textbase" (427-429). For our project, some of the comparable issues we grappled with include tagging "space" and "satire." More specifically, the kitchen or house becomes a domestic space that is turned upside down by the father's presence and/or the mother's absence therein: what is the best phrase to capture this phenomenon? And the use of satire, specifically relating to men dressed in women's clothes are, at the surface level, merely cross-dressing; yet, whether or not the artist and/or publisher intended, these images also evoked transgressive ideas about parenting regardless of which gender is performing the work. Our team collaborates on definitions of tags with the understanding that such definitions directly influence the search results returned by the user of our archive and on the exportation of Omeka metadata to other digital tools. The preliminary results from this digital humanities approach have revealed the degree to which representations of fatherhood and masculinity were central to the construction of both the pro- and anti-suffrage debate. Though Susan B. Anthony's politicking resulted in her memorialization in written and visual culture as the premier suffragist of the nineteenth century, a DH approach to the analysis of fatherhood in pro- and anti-suffrage postcards reveals that an additional version of visual history is being told.

We propose that the locus for images of ineffective and bewildered fathering in popular culture can be found in these anti-suffrage postcards, creating a trope that continued in twentieth-century popular culture. The film Mary Poppins (1964), for example, with its ridicule of the woman suffragist, presents a family that is ineffective because both parents are inattentive to their children. Similarly, Three Men and a Baby (1987), Cheaper by the Dozen (2003), and

countless television commercials for baby products satirize clueless fathers; in Mrs. Doubtfire (1993) the father character must dress as a woman in order to be a "good father." Since demonizing feminized fatherhood or valorizing national fatherhood had little immediate use for reformers once suffrage was achieved, we tentatively suggest that there was, in addition to the trope of the ineffective father, a second shift. For ex-suffragists in the 1920s and 1930s, there was a shift toward claiming the foremothers of women's suffrage across various visual forms. As postcards waned in popularity, the now-respectable victory of women's suffrage was memorialized in various forms. Photographs, sketches, and sculptures of reform women including Lucretia Mott, Abigail Adams, and of course Anthony herself abounded – proclaiming that Abraham Lincoln no longer need be the only father of the nation, but that women could equally be noble mothers of the nation.
Omeka: https://thesuffragepostcardproject.omeka.net/about; Twitter: @Suff_Postcards

2. Mauro Carassai, "Future directions in distant reading the subject: strategies and challenges in the computer-enhanced analysis of hybrid languages in Caribbean fiction"
In my paper I discuss the challenges hybrid languages can pose to the computer-enhanced textual analysis of topics such as the theme of human and machinic subjectivity in literary representations. As a work of speculative fiction that probes the limits of disciplinary divisions between digital studies and (New) American Studies, Nalo Hopkinson's Midnight Robber re-inscribes the universalizing post-humanist paradigm usually associated with technological culture within a linguistic creolization process that invests various sentient creatures (humans and machines). In the futuristic world described in the novel, Hopkinson creates a new language, a version of pidgin English that includes elements of Jamaican, Trinidadian, and Guyanese. One of the interesting aspects of this creolized English is that, from the grammar point of view, subjective pronouns indicating the direct object (i.e. the element of the sentence that receives the action performed by the subject) are frequently morphologically reconfigured in the subjective form. In other words, both for humans and machinic characters the English word "me" regularly appears as "I" and the word "us" regularly appears as "we" regardless of their grammar function as either subject or direct object.
Different textual analysis software (Voyant, Lexos, NLT) can help in different ways in the analysis of the theme of subjectivity by offering differing strategies to operate distinctions among the various pronouns. Some software can operate in terms of both "concomitant variations" and "cluster analysis" patterns when the goal is to differentiate human characters from machinic characters. However, things become more complicated when the goal is to analyze the different human and machinic "constructs" that Hopkinson builds in the novel. The fictional digital entities created by the author represent an array of intermediate cases both in terms of blending the human and the machinic and in terms of exhibiting properties of "relational artifacts" in Sherry Turkle's terms. As digital humanist John Burrows remarks, "in its traditional forms, textual analysis has to do with separating, distinguishing, and the like and it usually treats of single works." However, the forms of analysis I have tried to carry on in Hopkinson's book, often aims at gathering, combining and classifying – a practice that in Digital Humanities is usually associated with computational stylistics (stylometry) – where the application of software like Signature, stylo, of Stylene would be more fruitful.
Here is where the problem of the future direction of software useful to Digital Humanities research – and specifically for digital literary studies – interject the crucial problem of rethinking assumptions about textuality that have become established notion in structuralist and post-structuralist literary theory. To what extent is quantitative "style" representative of a specific authorial subject/speaker regardless of its ontological status as an existing author or a fictional character? To what extent is quantitative analysis software taking into account the conceptual transformations of our notions of subjectivity when subjects materializes identity along the ambiguous borders of inter-cultural linguistic domains? My paper ends with a few suggestions for incorporating software functionalities that might face future transformation of the notion of subjectivity as it relates to language use.

3. Laura Michelle Hale, "Sports data revolution: A case study in closing the disability sport data gap as a tool for improving sporting performance and increasing inclusiveness"

As sports has moved to take greater advantage of the large volume of statistical data available to do more complex analysis in the name of performance improvement, gaining competitive advantage and talent identification, large gaps have emerged primarily in the area of women's sport and disability sport. These gaps are a result of lack of interest, lack of commercial value in the data, and difficulties in accessing and storing these data. Consequently, women's sport and disability sport are put at a further competitive funding and commercialization disadvantage because efforts are concentrated in more the more lucrative men's sports.

For disability sports dependent upon performance data analysis to ensure fair play through continual analysis of classification systems, this gap can result in a situation where certain groups are underrepresented or put at a competitive disadvantage. This is particularly true in certain areas where people with different disability types such as cerebral palsy compete against people with amputations and spinal cord injuries. The nature of their disability is such that muscle and strength performance mirror the latter groups, but less well understood is the impact of cardiovascular endurance respiratory capacity. This can lead to people with cerebral palsy posting inferior performances, resulting in further exclusion from sport across all levels of sport. Large datasets including publicly available datasets that describe general disability characteristics against sporting performance based on classification and gender can provide greater insight, as opposed to the use of smaller scale data analysis that tends to be based around single events.

My paper explores the practical issues of creating a large-scale database for disability sport, through a project being worked on to create a database of disability sports information. The primary focus is on discussing data accessibility, the decision-making process for the software used, questions around how accessible data should be, and balancing the desire for openness to improve the lives of people with disabilities and improving their sporting performance with trying to monetize these efforts to enable the project to continue. Secondary to this, the paper explores some of the issues related to the construction of a large-scale sport database, that covers over fifty different sports for people from ten different disability groups with information dating back to the 1920s.

Biography

Kristin Allukian (University of South Florida)

Kristin Allukian completed her B.A. at Mount Holyoke College, her M.A. at Trinity College in Hartford, and her Ph.D. at the University of Florida. She spent two years as a Brittain Postdoctoral Fellow in Digital Pedagogy at the Georgia Institute of Technology. Currently, she is an Assistant Professor of English and Affiliate Faculty of Women's and Gender Studies at the University of South Florida. Her primary research area is nineteenth-century American women's literature. Her current book project uses feminist historical context and archival research to analyze representations of working women and labor systems in American women's literature and culture between 1839 and 1915. Her research and teaching interests include American literature before 1900; women's literature; and feminist digital humanities. Her work has appeared in academic journals including *Symbiosis, Women's Studies*, and *Ada: A Journal of Gender, New Media, and Technology*.

Mauro Carassai (California State University, Northridge)

Mauro Carassai is Assistant Professor of Liberal Studies at California State University Northridge where he teaches courses in Digital Humanities, Literary Theory, and American literature. He was a Brittain Postdoctoral Fellow at Georgia Institute of Technology in 2014-15 and a visiting Fulbright at Brown University in 2007-2008. His research combines literary theory, philosophy of language, and digital literatures within the larger frame of American literatures and American studies. His scholarly work has been published in journals such as *Culture Machine, LEA Almanac*, and *ADA – A Journal of Gender Media and Technology*. He co-edited a double issue for the *Digital Humanities Quarterly* titled "Futures of Digital Studies" and he is currently at work on a manuscript exploring problems and perspective in configuring an *Ordinary Digital Philosophy*.

Laura Hale (ParaSport News, Spain)

Laura Michelle Hale completed her Bachelors of General Studies and her Masters of Science in Instructional Technology at Northern Illinois University, and completed her PhD in Communications at the University of Canberra. She was the first person to serve as a Wikipedian in Residence at a sporting organization, serving in this role for both the Australian Paralympic Committee and the Spanish Paralympic Committee. Her primary research interests are social media research methodologies, and in women and disability sports. She has presented at the 2012 Australia Cycling Tourism Conference, World of Football II, Melbourne titled, "Filling an Information Void: Using Wikipedia to Document the State of and Promote Women's Soccer in Africa", and presented at a conference at the Australian Institute of Sport about Wikipedia. As an accredited journalist affiliated with the Wikimedia movement, she has covered three Paralympic Games, the 2012 Rollers and Gliders World Challenge, the 2012 IPC NorAm Cup, the 2013 IPC Alpine Skiing World Championships and a few other events.

Macroscopic Exploration of Large Text and Image Collections via Similarity Heatmaps

Peter Broadwell, Tomoko Bialock (UCLA Library), Hiroyuki Ikuura (Waseda University)

In the digital humanities, the term “macroscope” has come to denote a suite of interchangeable tools that contribute to all levels of humanistic inquiry, from close to distant reading. The work described here is part of an ongoing collaboration between faculty, researchers, and library staff — which we call HYU:MA (ヒュー:マ) — to realize the potential of the humanities macroscope in the uniquely challenging but also highly rewarding context of Japanese digital scholarship.
For this study, we adopt analytical and visualization approaches commonly used in the biological and behavioral sciences — specifically, similarity heatmap plots — to build prototype tools for comparing large digitized collections of Japanese texts and images. The ability to work with both texts and images is especially useful for Japanese cultural studies; accordingly, our research demonstrates the applicability of heatmap analysis to imperial anthologies of *waka* poems as well as to Edo-period theatrical illustrations.

Similarity heatmaps and the humanities macroscope

Comparison is one of the primary modes of humanistic inquiry. We propose that similarity heatmap visualizations (see Figure 1) are well suited to “macroscopic” comparative explorations of large collections of digital objects. A heatmap plot visualizes the relative degrees of similarity of each item in a collection to every other item; the items are enumerated along the vertical (Y) and horizontal (X) axes, and the color at point (X,Y) indicates the degree of similarity between items X and Y. The visualization can support any desired similarity metric; for text collections these may range from purely stylistic aspects such as sentence length to content-oriented measures like topic similarity. Crucially, our heatmap software features enable one to “zoom in” to a pairwise comparison between any two items in order to investigate the precise nature of their similarities and differences. Recent advances in open-source visualization software — particularly the Plotly data-science suite — facilitated the development of these features.

Text corpora: *waka* anthologies

We applied our comparison heatmap analysis to the *Nijūichidaishū*, the 21 imperial *waka* poetry anthologies compiled from 905 through 1439. The parallel construction of the anthologies — following the progression of seasonal topics and other genres established in the earliest collections — provides rich opportunities for macroscopic analysis via the heatmap interface. Indeed, after ordering the poems sequentially and applying word- and topic-based similarity metrics, we noted intriguing nuances of these parallel structures in the visualization, which we investigated via the heatmap’s interactive “zoom-in” features (see Figures 1-3). Our scholarly collaborators suggested that certain coloration patterns appearing over large portions of the map likely correlate with significant literary-historical trends, while other features are more difficult to explain and present opportunities for new discoveries.

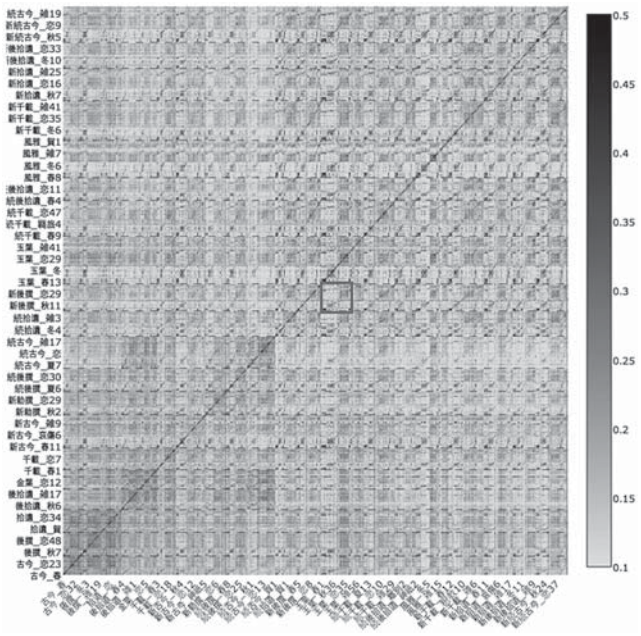


Figure 1: A similarity heatmap of the texts of the *Nijūichidaishū*, with groups of poems labeled by their anthology name and genre, e.g., 古今 _ 賀 .[1] The area highlighted in red is shown in detail in Figure 2. Horizontal dashed lines show boundaries between the anthologies on the left axis. Yellow shading indicates low similarity between texts, while blue indicates high similarity, based on a TF-IDF weighted n-gram cosine similarity score from 0 (least similar) to 1 (most similar). The color range has been tuned to highlight similarities. The dark blue "threads" running parallel to the main diagonal indicate parallelisms in the content of the poems in different anthologies.

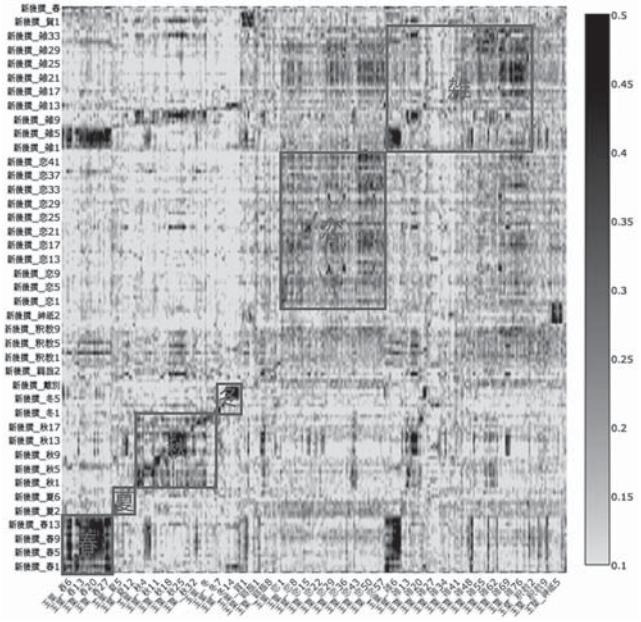


Figure 2: A zoomed-in comparison heatmap of the poems of the *Shin Gosen Wakashū* (新後撰和歌集, left axis) and the *Gyokuyō Wakashū* (玉葉和歌集, bottom axis) anthologies, with corresponding genre sections highlighted. The thickness of the diagonal "threads" indicates the degrees to which the anthologies follow the same time-based topical progressions (early autumn to late autumn, etc.) within corresponding seasonal genres. The comparatively large and varied sections for "love" (恋) and "miscellaneous" (雑) poems, by contrast, exhibit fewer internal correspondences across anthologies.

[1] The sequence numbers following the genres are unique to this visualization and should be ignored.

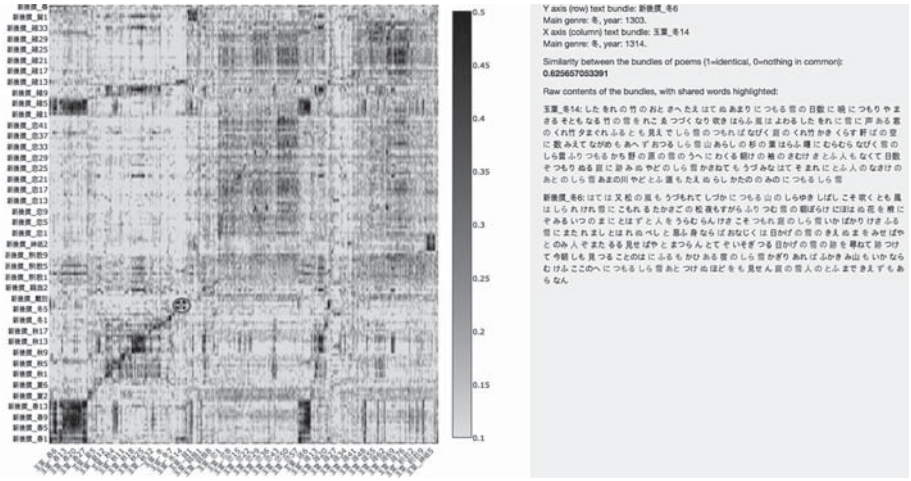


Figure 3: A "zoomed-in" comparison of the texts of 10 poems from each of the two anthologies shown in Figure 2.

Image corpora: *ezukushi banzuke*

The heatmap techniques described above for text corpora also are applicable to large image collections, which is exciting given the prominence of visual art throughout much of Japanese history and the availability of large digitized image collections online. In the future, it may even be possible to automate similarity analyses and heatmap generation comparing multiple image corpora if the collections' metadata and derivative "thumbnail" images have been described and made accessible via the Internet Image Interoperability Framework (IIIF) standard.

Software to find and track reuse of visual figures works well for print-based materials (see Figure 4), while emergent deep learning-based approaches allow for comparisons that take into account a broad range of stylistic factors (see Figure 5), rather than verbatim figure reuse. Though such work remains experimental, the ability to track stylistic features over time and even across media may contribute important new insights to digital art history and media studies.

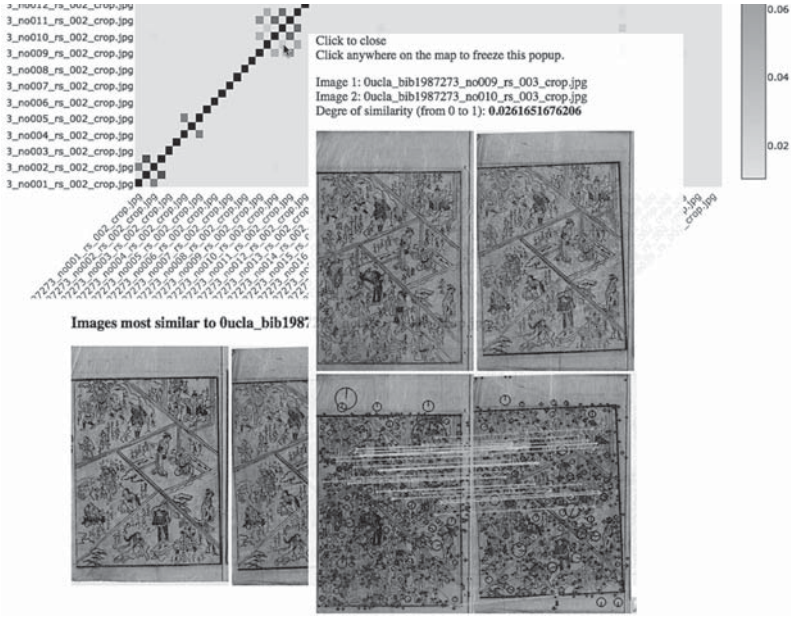


Figure 4: Detail view of the comparison heatmap for a set of kabuki playbills, highlighting pages from two different playbills for which the upper portion of the original print was reused. Shared visual "keypoints" (also known as visual "words") have been detected with a scale-invariant feature transform (SIFT) and matched via an approximate nearest neighbor algorithm, with the similarity values for the heatmap based upon the number of close matches between each pair of images.

The materials shown here are a collection of 505 page images of playbills from kabuki theaters in Osaka (*ezukushi banzuke*) from the 1780s to 1870s. As with texts (see above), interactive similarity heatmaps of image sets like these provide a comprehensive view of the pairwise similarities across entire corpora, in contrast to scatter plots, clustering diagrams, and query-based search interfaces, which emphasize some relationships while obscuring others. The ability to group and reorder the items compared on the heatmap and to "zoom in" to the comparison of any two items also greatly facilitates the "macroscopic" exploration and study of a corpus.

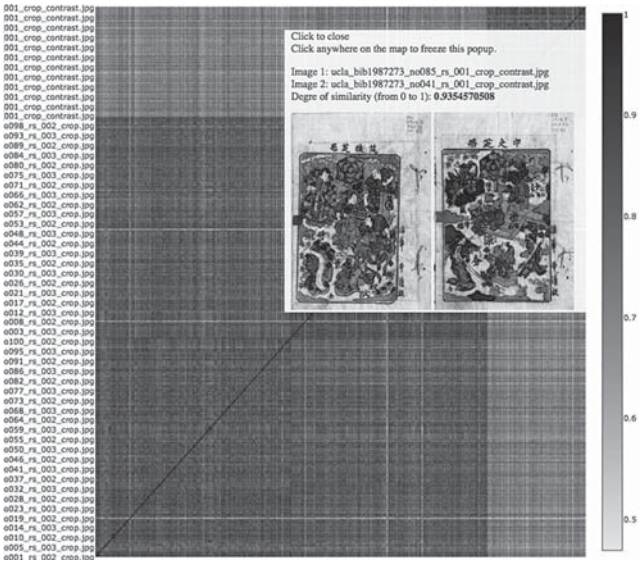


Figure 5: A similarity heatmap for the same set of kabuki playbills from Figure 4, based upon visual features extracted with the TensorFlow neural network library. Overall image similarities were computed as the cosine similarities of the features of the "bottleneck" layer, as determined by the Inception neural network. Playbill covers are compared to each other in the top right section of the heatmap. The highlighted pair of covers has few visual "words" in common (as in Figure 4), but the neural network nevertheless identifies these images as being highly similar stylistically.

A Neural Network Approach to Historic Linguistic Change

Eun Seo Jo, Dai Shen, Michael Xing, Mark Algee-Hewitt (Stanford University)

Abstract

Historians have traditionally only used qualitative methods to identify historic moments of linguistic change in discourse but this is limiting in objectivity and volume. Numeric representations of language allow us to quantify and compare the significance of discursive changes and thus capture linguistic relations over time. We compare deep learning methods of quantitatively locating when the usage of words in context changes. We introduce the method of charting the moving average of perplexity in a language model trained on chronologically sorted data. These results correspond to vectorized language we derive from a classification model. We apply our models on a historical diplomatic corpus to find results showing that the 1910s and 1940s in American diplomacy proved to be notable moments in linguistic change. With this example we hope to introduce applications of deep learning methods to Digital Humanities usages.

Introduction & Literature Review

Humanists appreciate that linguistic change over time can reflect critical shifts in historic context whether they be in discourse, ideology, or access to communications technology[3][4]. Historians have relied on context to narrate the changing meaning of words and language. George Chauncey shows how the tense Cold War atmosphere and the end of the Prohibition instigated a narrower usage of terms such as ”gay” and change in gender classification from cultural attributes to choice of sexual object [1]. Literary scholar Fliegelman argues that the Declaration of Independence was written in a style that highlights oratorical impact including rhythmic pauses and stresses [2].

This work proposes a deep learning approach to identifying moments of major deviations in language. Traditional humanities works trace the meaning of words over time across multiple archives through close-reading, such as those above.

Works in computational linguistics have introduced quantitative methods, the most relevant of these is Juola’s contribution which uses KL-divergence measures between pairs of documents of differing year gaps at various time periods [4]. Juola shows that the 1950s were times of significant linguistic change attributable to post-WWII culture and developments in television technology. Juola’s work has two main shortcomings. The first is that KL-divergence alone does not give a qualitative, historical explanation as to why there were changes and the author defers to postulations of notable contextual events. Second, while we are able to quantify relative linguistic change by time periods we cannot tease out single words to track their individual evolution over time. Hamilton et.al’s recent work addresses this limitation with word vectors. They generate word embeddings from PPMI, SVD, and word2vec delineated by time period and compare these embedding matrices across time by taking their orthogonal Procrustes solutions, derived from an SVD application (except the PPMI matrix)[5]. Using these results, they propose several laws of linguistic evolution and most importantly showcase how individual words can be traced over time by comparing vectors from different eras[5]. Linguistic change has been a topic of interest for the Digital Humanities community. Ted Underwood, for instance, uses LDA topic modeling on a time segmented corpus to visualize the changing prevalence of generated topics [6]. He charts the changing proportion of a topic of words through time as a topical insight to changing language over time.

Our main intervention is the use of deep learning to identify periods of dramatic linguistic change at the token-level. Similar to Juola’s work, we are also examining language as a whole, though in future work we plan to trace the meaning of individual words. One reason why deep learning approaches have been successful in NLP tasks is because the use of word vectors that

encode multi-dimensional meaning of words based on co-occurrence with context words. Because we initialize our model with pre-trained vectors, we utilize word meaning in our generative RNN, giving higher overall accuracy. With word vectors, the perplexity scores reflect not just the sequence of words but also the meaning of them. Another advantage of neural nets over Juola's KL-divergence methodology is computational convenience - that we can build the model to run the entire chronologically ordered corpus through to compare the relative perplexity scores across time, whereas Juola's method requires each era to be computed and tested separately. Further, this paper also hopes to introduce a relatively under-utilized tool in Digital Humanities – neural nets, by way of an example.

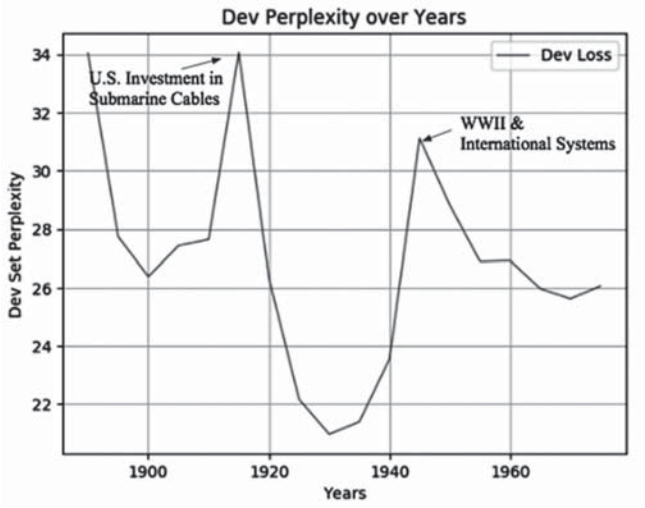


Figure 1: Development set average perplexity over the five-year eras. The perplexity peaks at two places through the years, 1915 and 1945, roughly corresponding to the start of WWI and the end of WWII.

Our main approach trains an RNN language model over temporally sequential sets of training data. We observe the changes in perplexity over these eras and find peaks at moments of dramatic language shifts. We also test an RNN classification model that predicts the year bucket of a given document. Here we use the softmax weights to demonstrate topical and semantic change over historical eras. The example corpus is a set of diplomatic texts containing historical communiques.

Dataset

We collected 238,097 historical diplomatic documents that span from 1860 to 1983. We performed stratified sampling to overcome the significant variance in frequency of documents over the year buckets. We also capped the number of documents at 5000 to smooth out the disproportionate representation of certain years. For both models, we use GloVe vectors with 400,000 vocab size of 100 dimensions.

Experiments

Document Classification

Recurrent classification model with GRU cells

We then trained and tested on an RNN classification model to classify documents into their corresponding year buckets. The RNN classification model consists of a single layer of GRU cells and prediction is made from the output of the last cell. Each GRU cell has the following composition:

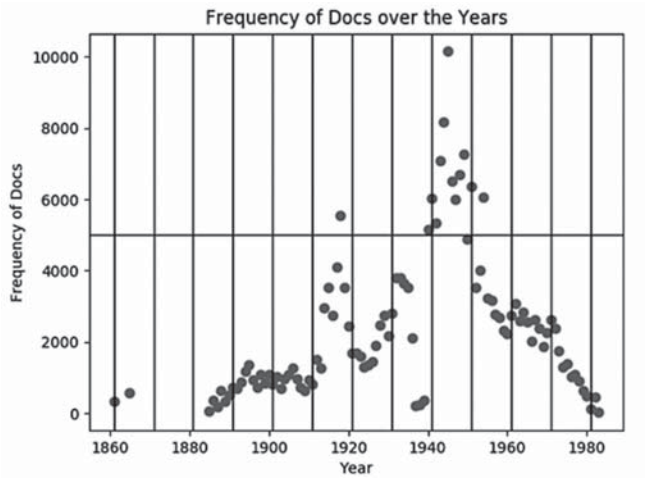


Figure 2: Frequency of Documents over the Years. Documents were capped at 5000 per year.

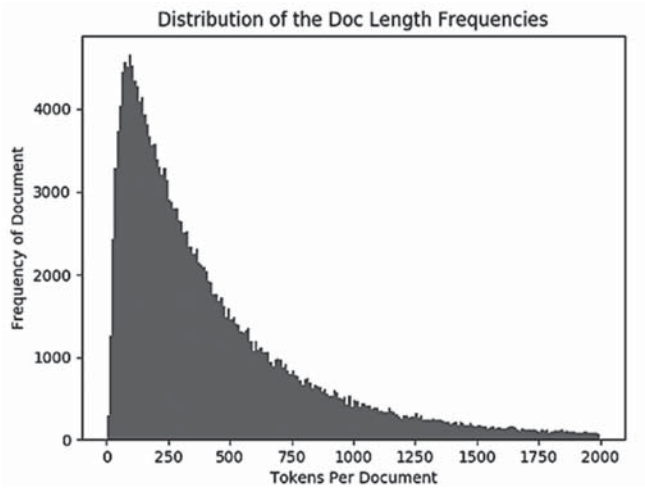


Figure 3: Histogram of the documents by token count. We chose to truncate all documents at length 600 and pad those shorter than.

$$z^t = \sigma\left(x^t U_z + h^{(t-1)} W_z + b_z\right) \quad (1)$$

$$r^t = \sigma\left(x^t U_r + h^{(t-1)} W_r + b_r\right) \quad (2)$$

$$o^t = \tanh\left(x^t U_o + \left(r^t \otimes^{(t-1)}\right) W_o + b_o\right) \quad (3)$$

$$h^t = z^t \otimes h^{(t-1)} + (1 - z^t) \otimes o^t \quad (4)$$

$$(5)$$

Following the last time-step we make the year prediction as follows:

$$\text{pred} = \text{softmax}(h^{(last)} U + b_2) \quad (6)$$

where *pred* is the probability distribution of the document being in each decade. We applied dropout to prevent overfitting. We expect GRUs to have less of a problem with vanishing gradient and learn longer sequence data. To prevent exploding gradient, we integrate gradient clipping. We expect this model to better capture sentence structures than the plain bag of words (BOW) baseline.

Language Model

Our second model is a language model that trains by predicting the next word given the current and all previous words. We use the same RNN schematics as the previous model but make a prediction at each time step. The output of each GRU cell is a probability distribution over 400,000 classes, our vocabulary size. The GRU composition changes to the following:

$$z^t = \sigma\left(x^t U_z + h^{(t-1)} W_z + b_z\right) \quad (7)$$

$$r^t = \sigma\left(x^t U_r + h^{(t-1)} W_r + b_r\right) \quad (8)$$

$$o^t = \tanh\left(x^t U_o + \left(r^t \otimes h^{(t-1)}\right) W_o + b_o\right) \quad (9)$$

$$h^t = z^t \otimes h^{(t-1)} + (1 - z^t) \otimes o^t \quad (10)$$

$$\text{pred} = \text{softmax}\left(h^{(t)} U + b_2\right) \quad (11)$$

$$(12)$$

We train our language model data in chronological order so that it is able to learn in sequence. We divided them into half decade buckets. We make the first 85% of each decade as the training data and the last 15% of each decade as the development data (see figure 4). We train(85%)-dev(15%) in sequence over the eras.

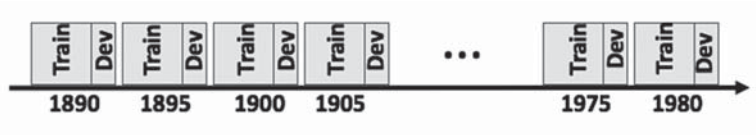


Figure 4: Diagram of Language Model Document Feeder: The 5 year corpora are fed into the model in sequence, starting from 1890.

Results

RNN Classification

With 30 epochs of training the RNN classification model, we were able to achieve f1 of over 80% as shown in figure 7. Figure 5 and 6 show the loss and accuracy respectively of the RNN compared to the baseline model. The RNN outperforms the bag of words baseline consistently throughout the epochs and reaches the peak performance after fewer epochs. This demonstrates that sentence structure played a significant role in predicting the year of a given document. We can also see the results of the RNN trained with the stop-words only corpus. The accuracy performs surprisingly well, surpassing 50%, indicating that stylistic changes were significant enough to be captured.

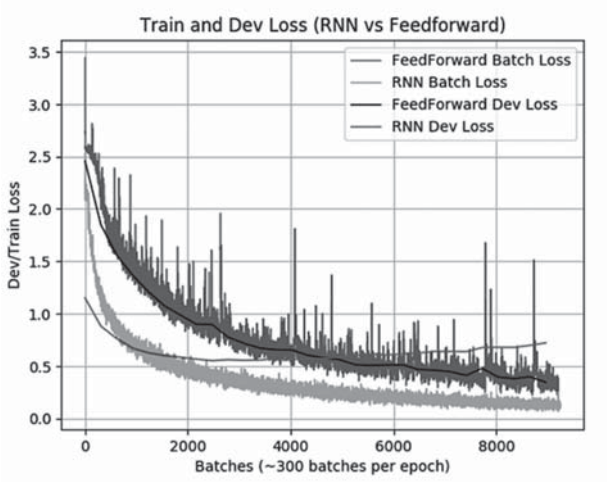


Figure 5: Comparison of Dev set loss over 30 epochs for RNN and Feedforward models

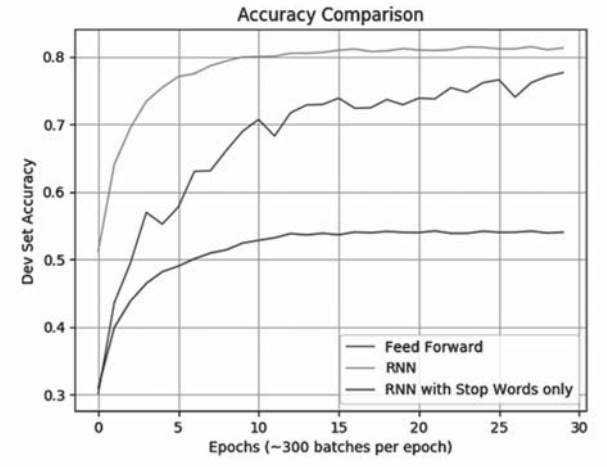


Figure 6: Comparison of dev set accuracy over 30 epochs for RNN, RNN with only stop words, and Feedforward models

F1 Scores of Models	
Model	F1 Score
Feedforward	72.0
RNN	80.3
RNN-stop words	50.2

Figure 7: Test F1 Scores of RNN, Feedforward

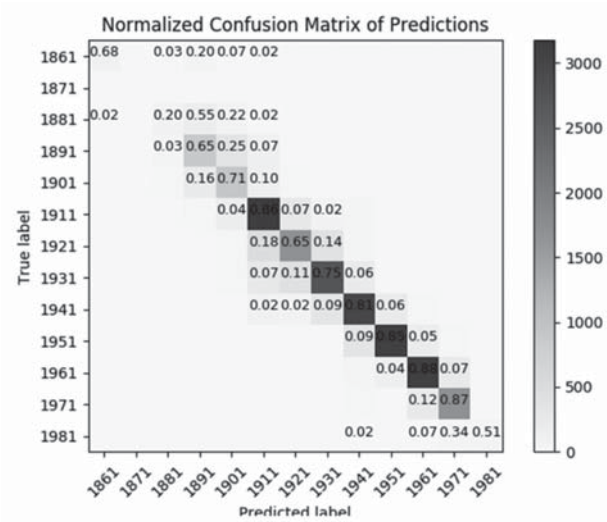


Figure 8: Confusion matrix of the feed- forward baseline model's prediction of year classes after 30 epochs. The first class corresponds to the decade [1861,1870] and so on.

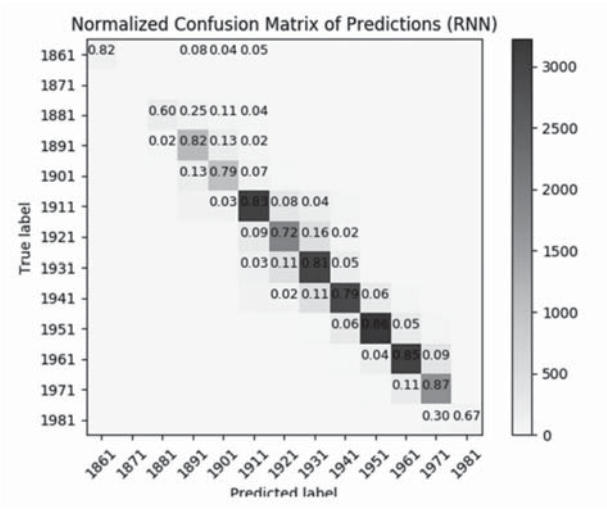


Figure 9: Confusion matrix of the RNN models' prediction of year classes after 30 epochs

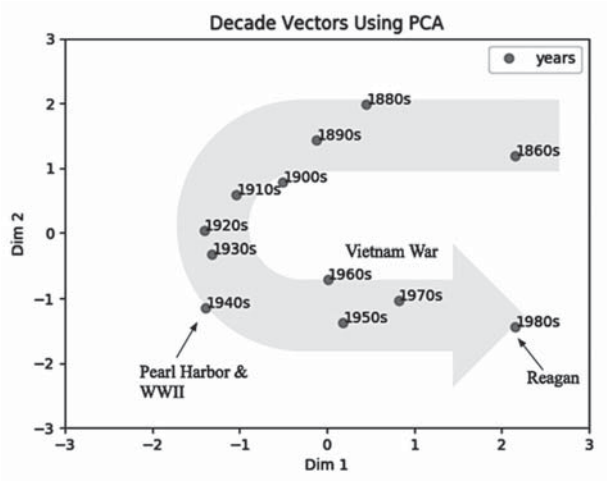


Figure 10: PCA visualization of the decade classes from the softmax weights. Dimensions reduced from 300 to 2. Note: 1870 is not graphed because the class bucket was missing data.

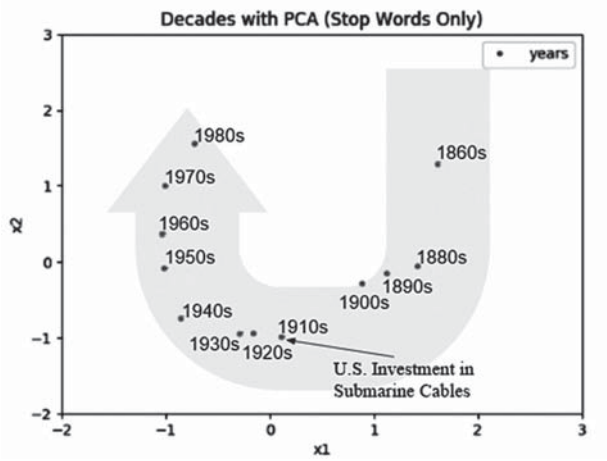


Figure 11: PCA visualization of decade classes from the softmax weights trained with non stop words and punctuation . Dimensions reduced from 300 to 2. Note: 1870 is not graphed because the class bucket was missing data.

To visualize the historical trends in language, we took the softmax weights, the *U* matrix of dimension [class × hidden size] in:

$$\text{pred} = \text{softmax}(h^{(t)}U + b_2)$$

and reduced the dimensions using PCA and charted the decade vectors in 10 and with only stop words in 11. The visualizations show overall trends in language through the century and a half. Figure 10 shows how general language, including ideologies and issues of concern, changed over time. Figure 11 maps the changes in purely stop word usages over time, an attempt to focus on stylistic change. It is notable that while 10 shows greater variance in changes from decade to decade, 11 appears more consistent, results which verify our expectations about topics and style. Given that the dataset is a collection of diplomatic documents, we labeled several notable historical events for interpretation. The distance between 1940 and 1950 in 10 is worth highlighting as post-WWII and the commencement of international organizations such as the UN(1945), WHO(1948), and IMF(1945) changed American diplomatic language to incorporate more vocabulary concerning global procedures. In figure 11 the jump in style from 1900 to 1910 aligns with historical literature on diplomatic language. The consensus is that eyeing the geopolitical and strategic potential of expanding the reach of telecommunications coverage, the U.S. State Department made significant investments in submarine cables during the first world war, which altered the constraints of telegraphic expression [4].

RNN Language Model

The results of the RNN Language Model corroborate our results from the RNN classification model. Figure 1 charts the changing dev set perplexity over the years at five-year intervals. If language is consistent enough throughout the years we would expect the perplexity to fall consistently as we feed more data. However, the results show perplexity spiking at certain historical moments, in 1915 and 1945, indicating the change in language was dramatic enough to confuse the model trained on the previous years' language. These two moments of spike align with the our results from above.

Conclusions

The two RNN-based approaches ultimately produce similar results, showing that the 1910s and 1940s were moments of significant linguistic change in American diplomacy. Our experiments show that RNN models can be useful tools for measuring linguistic and topical change over time. Finally, there are still many avenues to expand on the task of linguistic evolution. One direction would be to engineer creative methods of interpreting the features of linguistic change.

References

[1] George Chauncey. Gay New York: Gender, Urban Culture, and the Making of the Gay Male World 1890-1940. BasicBooks, New York, NY, 1994.
[2] Jay Fliegelman. *Declaring Independence: Jefferson, Natural Language, & the Culture of Performance*. Stanford University Press, Stanford, CA, 1993.
[3] David Paull Nickles. *Under the Wire: How the Telegraph Changed Diplomacy*. Harvard University Press, MA, 2003.
[4] Juola, P. Computers and the Humanities (2003) 37: 77. doi:10.1023/A:1021839220474
[5] Hamilton, William L., Jure Leskovec, and Dan Jurafsky. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change", 2016, https://arxiv.org/pdf/1605.09096.pdf.
[6] Ted Underwood and Andrew Goldstone. "What can topic models of PMLA teach us about the history of literary scholarship?", 2012, https://tedunderwood.com/2012/12/14/what-can-topic-models-of-pmla- teach-us-about-the-history-of-literary-scholarship

Vector based measure of semantic shifts across different cultural corpora as a proxy to comparative history of ideas

Alexandre Gefen (Université Paris Sorbonne), Mark Andrew Algee-Hewitt, David McClure (Stanford University), Frédéric Glorieux, Marianne Reboul (Université Paris Sorbonne), J.D. Porter (Stanford University), Marine Riguet (Université Paris Sorbonne)

Our joint research group, a collaboration between Paris Sorbonne and Stanford University, traces the parallel birth and evolution of key modern literary and aesthetic concepts in French and English using vector based semantic analysis methods. This inter-institutional, multilingual, collaboration aims not only to explore the development of “literature” as a distinct discursive and disciplinary field, but also to discover the mutual influence — or, perhaps, surprising independence — of these two closely related national literary traditions. By sharing our methods and results, we hope to foster the use of distributional semantics as a useful tool for investigating the comparative history of ideas in the literary field.

For this project, we have elected to compare two corpora of important and representative XIXth century magazines: in France, La Revue des Deux mondes (1829-1893), and in English, the Blackwood's Magazine (1817-1880), both influenced by the emergence of Romanticism. We started with a simple word, « littérature/literature » (9100 occurrences in the Blackwood corpus, 12400 occurrences in the Revues des deux mondes) — a word that we know has grown specialized during the century — in order to plot its semantic evolution and semantic space.

In this paper, we discuss the methods we have developed to:
Select, prepare and compare dissimilar corpora of old newspapers. After many discussions, we identified these two magazines as the most equivalent single-periodical literary corpora in our respective national literatures. To maintain this equivalence, we restricted the two corpora to the same time period, 1830-1880. We preprocessed both corpora with custom methods conversion to TEI encoding, and lemmatized them with Alix (Alix is a custom-designed lemmatizer and Part-Of-Speech tagger) in French, with the Stanford CoreNLP tagger in English.

Graph the comparative historical evolution of our key word « littérature/literature » in vector space using word2vec and Glove vectors. This builds upon our previous work with the HathiTrust corpus in which we identified the lexemes with the most distinctive histories in relation to literature. In this pilot attempt, we identified words that appeared frequently on the same page as “literature” but with significant variation over time -- words that shifted from low to high levels of correlation with “literature” across literary history, and vice versa. It is our contention that by moving from this initial method to a vector-based analysis of our selected periodical corpora, we will better be able to capture the nuances of the semantic field of “literature” over time.

Define a reliable way to measure polysemy. First, we build a set of chosen concept words with their equivalent in both languages in WordNet. Eg. literature:littérature ; poetry:poésie ; art:art ; etc. Next, we build a set of bigrams for each of the words of this set, with the ten most associated adjectives. We then compute the cosine distance between each bigram and apply the methods used in Köper and Schulte im Walde 2014: once the cosine distance is computed for each bigram vector, build a matrix with ranking proximity of those vectors, and compute their mean rank. We are then able to determine a score that would weight the computing of the clustering coefficient of bigrams. As a low-rank bigram should in our assumption represent a highly polysemic term, the lower the rank, the higher the weight on the clustering coefficient. We then determine the clustering coefficient of each term in diachronic sequences, to draw a dynamic representation in Gephi.

This method allows us to find possible semantic attractors in the evolution of key literary concepts and to plot the behaviour of specific semantic word pairs (for instance, littérature/poésie and literature/poetry). In our final cultural interpretation, we compare semantic matrices and historical evolutions in the two corpora/languages and confront the results with known existing hypotheses in the history of aesthetic concepts in order to confirm the supposed specialization of the field of literary studies. By focusing on concepts that undergo dramatic transitions during the period of study, our analysis brings together the study of diachronic conceptual change and synchronic polysemy, allowing us to probe how multiple senses of a word coexisting in tension can eventually give rise to changes in meaning.

Bibliography

[1] Boleda, Gemma, Sebastian Padó and Jason Utt. “Regular polysemy: A distributional model.” SemEval 12 (2012): 151-160.
[2] Hamilton, William, Jure Leskovec, J. and Dan Jurafsky. “Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change.” arXiv:1605.09096 [cs.CL] (2016)
[3] Jorge-Botana, Guillermo, José A. León and Ricardo Olmos. “The representation of polysemy through vectors: some building blocks for constructing models and applications with LSA.” Int. J. Cont. Engineering 21 (2011): 328 – 342.
[4] Köper, Maximilian, and Sabine Schulte im Walde. "A Rank-based Distance Measure to Detect Polysemy and to Determine Salient Vector-Space Features for German Prepositions." LREC 9 (2014): 4459-4466.
[5] Mu, Jiaqi, Suma Bhat, Pramod Viswanath. “Geometry of Polysemy.” arXiv:1610.07569 [cs.CL] (2016).
[6] Ravin, Yael, and Claudia Leacock (editors). Polysemy: Theoretical and Computational Approaches. New York: Oxford UP, 2000.
[7] Springorum, Sylvia, Sabine Shulte im Walde and Jason Utt. “Detecting Polysemy in Hard and Soft Cluster Analyses of German Preposition Vector Spaces.” IJCNLP 6 (2013): 632-640.

FAIRness for Citizens: Workflow and Platform for Open Data with a Case Study on Edo Cooking Recipes

Asanobu Kitamoto
(Research Organization of Information and Systems / National Institute of Informatics)

FAIRness in triadic co-creation

Data-driven science, or digital humanities, has three key stakeholders, namely researchers, machines and citizens, where three groups have complementary roles to cooperate, or what we call triadic co-creation. Researchers deepen the body of knowledge by their expertise, machines increase the size of knowledge by their computing power, and citizens enhance the diversity of knowledge by their wide interest. Open data is a critical ingredient shared by all stakeholders, but what is the best practice for promoting the usage of data by each stakeholder?

We suggest that a starting point to consider this issue is FAIR guiding principles [1], which was proposed by FORCE 11 (the Future of Research Communications and e- Scholarship), a community of scholars, librarians, archivists, publishers and research funders. FAIR stands for Findable, Accessible, Interoperable and Reusable, and the guiding principles pay special attention to "FAIRness for machines" to support automated analysis by computing agents. In a triadic co-creation environment, however, the challenge is to generalize this concept to "FAIRness for researchers" and "FAIRness for citizens" to incentivize the usage of data by all groups of stakeholders.

As an example, in November 2016, we released open datasets from Center for Open Data in the Humanities (CODH) of Research Organization of Information and Systems (ROIS) in collaboration with National Institute of Japanese Literature (NIJL) [2]. It consists of three datasets as follows.

i. Dataset of Pre-modern Japanese Text (PMJT) [3]: digital images of 701 pre- modern Japanese books fully digitized from the top to the bottom. Useful for reading and downloading the original books.
ii. Dataset of PMJT Character Shapes [4]: about 403,242 records of Unicode code point and XYWH coordinates on PMJT images. Useful for training optical character recognition (OCR) software.
iii. Dataset of Edo Cooking Recipes [5]: about 100 recipes of egg dishes from the book published around 1785. Useful for cooking.

Roughly speaking, (1), (2) and (3) datasets are targeting researchers, machines and citizens respectively.

Each dataset should be designed to satisfy its own FAIRness criteria, but the biggest barrier to realize FAIRness for non-researchers is the language and characters. Text of old Japanese characters is readable only for researchers, and not for machines nor for citizens. For machines, the lack of optical character recognition (OCR) software and manually transcribed text is a barrier for automated machine learning-based analysis. For citizens, the lack of fluent readers, which are estimated to be only a very small fraction of native Japanese speakers, is a barrier for understanding knowledge in the books. This means that the open datasets are literary not citizen-readable data.

Workflow and Platform for FAIRness

As introduced, our target dataset, "Dataset of Edo Cooking Recipes," is derived from a cooking book published in 1785 (Edo Period) with the title *Mambo Ryori Himitsubako* [6]. It consists of more than 100 cooking recipes of egg dishes, and our motivation is to let citizens use them in their daily lives and create derivative recipes to expand our knowledge about culinary culture in Japan.

To reduce the barrier for citizens, we designed a workflow to increase the FAIRness.

i . Organize digital images in the Dataset of PMJT.
ii . Transcribe text from digital images.
iii . Translate old Japanese text to modern Japanese text.
iv . Edit modern Japanese text into cooking recipes.

In the later stage within the workflow, the content of books is more human readable and accessible. In addition, the final step (iv) improves the interoperability of knowledge representation in terms of quantity, ingredients and tools.
In terms of quantity, a recipe in the original book lacks detailed descriptions about time and amount, probably because the book was published for professionals of the age. We however added estimated quantitative information not in the original books to make the recipe actionable for citizens by reducing both the risk of failure and the necessity of trial and error. In terms of ingredients, we changed some of old or local ingredients to similar ones that are easier to obtain now. In terms of tools, we replaced old cooking tools to modern ones, such as refrigerator and food processor, so that we can take advantage of today's civilization. The last step is to add photographs for each step, and the final photograph is well decorated for attractive impression. Photographs help people imagine the whole process of cooking and convince the reproducibility of the recipe.

Once the transformation of content is finished, where to deposit data is an important choice. Our choice was to use Cookpad [7], the largest cooking recipe service in Japan, because it is a platform which is already familiar with citizens in their daily lives. After depositing Edo Cooking Recipes into Cookpad, we received unexpectedly large response from citizens through various mass media, including television, radio, and magazines [8], and social media, including Twitter and Facebook.
We analyze that this success is because of better findability of the dataset in an interoperable and actionable format on the platform where large number of citizens is already familiar with. In addition, Cookpad has a function called "Tsuku-repo," which is designed to share photographs and captions about derivative recipes from the original one. To take advantage of this built-in function, we could also enhance the reusability of data.

Figure.1: Cookpad Edo Cooking Kitchen

POCiNP Approach

As introduced, the effective combination of workflow and platform improved the FAIRness of cooking recipe data for citizens. We call this approach as “Putting Old Content into New Platforms (POCiNP)” approach. When old books are digitized, their media are transformed from atoms to bits, but the platform of content remains to be the same, namely book forms. A new platform such as Cookpad requires the transformation of content because it is designed to offer a new form of knowledge. A surprising observation is the inversion of freshness; once old content was put into a new platform, citizens perceive old content as fresh content which was never seen on the new platform. We believe that the same approach can be applied to other combination of content on platforms, but this is left for future work.

Acknowledgment

The author would like to thank Mrs. Ui Ikeuchi at University of Tsukuba for personal communication that led to the original idea of the paper.

References

[1] Wilkinson, Mark D. et.al., The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data, Vol. 3, 2016. doi:10.1038/sdata.2016.18.
[2] Asanobu KITAMOTO, Kazuaki YAMAMOTO, "Construction of trans-disciplinary data platform that explores open data in the humanities", IPSJ SIG Computers and the Humanities Symposium 2016, pp. 117-124, 2016-12 (in Japanese)
[3] Dataset of Pre-modern Japanese Text (PMJT), http://codh.rois.ac.jp/pmjt/.
[4] Dataset of PMJT Character Shapes, http://codh.rois.ac.jp/char-shape/.
[5] Dataset of Edo Cooking Recipes, http://codh.rois.ac.jp/edo-cooking/.
[6] Mambo Ryori Himitsubako, 1785. doi:10.20730/200021712.
[7] Cookpad, https://cookpad.com/.
[8] CODH News, http://codh.rois.ac.jp/news/

FloraCultures: Conserving Plant-Based Cultural Heritage in Australia through Digital Approaches

**Paul Arthur (Edith Cowan University),
John Ryan (University of New England),
Heather Boyd (Edith Cowan University)**

Perth, Western Australia, is an urban area of high biodiversity. Botanist Stephen Hopper claims that "Perth is one of the world's most biodiverse cities, especially in relation to plants [···] The rate of discovery of new plants here, for example, is equivalent to the rate of discovery in many of the rainforests" (cited in Perth Biodiversity Project, n.d., p. 1). With the increasing loss of bushland to urban and suburban development, however, plant conservation faces a considerable challenge. Yet, in protecting biodiversity, cities at the same time gain opportunities to conserve the diverse forms of heritage, including cultural heritage, associated with their plants. This heritage involves plants as food, ornamentation, medicine, and fibre; as literary, artistic and historical objects; and as sources of community memory, cultural identity, and personal well-being.

FloraCultures develops theoretical and practical approaches to conserving botanical heritage in Western Australia. The project applies concepts of tangible and intangible heritage to plant conservation. Intangible heritage suggests the "forms of cultural heritage that lack physical manifestation. It also evokes that which is untouchable, such as knowledge, memories and feelings" (Stefano, Davis, and Corsane, 2012, p. 1). The 2003 UNESCO Convention for the Safeguarding of the Intangible Cultural Heritage articulates five manifestations of intangible heritage, including "oral traditions and expressions [···] social practices, rituals and festive events [···] knowledge and practices concerning nature" (UNESCO, 2003, p. 2).

"Knowledge and practices concerning nature" is of particular relevance to the project. Scientific knowledge of Western Australian plants is available online through the website FloraBase and through texts such as Flora of the Perth Region (Marchant, 1987) and Perth Plants (Barrett and Tay, 2016). Released by the Western Australian Herbarium in 1998, FloraBase is the "authoritative source" for plant knowledge in the state (Department of Environment and Conservation, 2017). A comparable resource for Western Australian plant-based cultural heritage is presently unavailable, and so FloraCultures aims to fill a significant gap. Notable examples of digital plant-based heritage do exist elsewhere. For instance, the Native American Ethnobotany Database, published by the University of Michigan, highlights plants as foods, fibres, and medicines in Indigenous American societies (Moerman, n.d.). For Western Australian species, some ethnobotanical information—interview materials and Indigenous names—is featured on the website Nidja Beeliar Boodjar Noonookury Nyininy or A Nyungar Interpretive History of the Use of Boodjar in the Vicinity of Murdoch University, Perth, Western Australia (Collard, 2003).

Drawing from current developments in heritage protection, memory studies, and digital design, this paper presents an overview of the project's model for conserving Perth's botanical heritage. The premise behind FloraCultures is that digital humanities approaches can render cultural information about Perth's plants accessible. The project consolidates materials dispersed widely across numerous traditions and sources, including Whadjuck Nyoongar (the Indigenous people of Perth) and colonial European expressions of knowledge and attachment to plants. Literature, art, historical accounts, and the memories of contemporary conservationists are brought together under the umbrella of the research. Thus, a precedent for the conservation of cultural, social, literary, historical, and artistic knowledge of Australian biodiversity is developed.

References

[1] Barrett, R. and E.P. Tay. (2016). Perth Plants: A Field Guide to the Bushland and Coastal Flora of Kings Park and Bold Park. Collingwood: CSIRO Publishing.
[2] Collard, L. (2003). “A Nyungar interpretive history of the use of boodjar in the vicinity of Murdoch University, Perth, Western Australia.” Retrieved May 7, 2017, from http://wwwmcc.murdoch.edu.au/multimedia/nyungar/menu9.htm.
[3] Department of Environment and Conservation. (2017). “FloraBase: The Western Australian flora.” Retrieved May 7, 2017, from http://florabase.dec.wa.gov.au/.
[4] Marchant, N. (1987). Flora of the Perth region. Perth: Western Australian Herbarium.
[5] Moerman, D. (n.d.). “Native American ethnobotany: A database of foods, drugs, dyes and fibers of Native American peoples, derived from plants.” Retrieved May 7, 2017, http://naeb.brit.org/.
[6] Perth Biodiversity Project. (n.d.). What is the Perth biodiversity project? West Perth: Perth Biodiversity Project. Retrieved May 7, 2017, http://www.lbp.walga.asn.au/Portals/1/Templates/docs/background_pbp.pdf.
[7] Stefano, M., Davis, P., and Corsane, G. (2012). Touching the intangible: An Introduction. In M. Stefano, P. Davis & G. Corsane (Eds.), Safeguarding Intangible Cultural Heritage (pp. 1-5). Woodbridge, England: Boydell & Brewer.
[8] UNESCO. (2003). Convention for the safeguarding of the intangible cultural heritage. Paris, France: UNESCO.

Making a Mark: Tradition and Technology in the Digitization of the Japanese Votive Slips Collection at the University of Oregon

Kevin McDowell (University of Oregon)

Abstract

During the late Edo period a craze developed among *nōsatsu* aficionados for commissioning the production of highly decorative, multi-colored exchange slips (*kōkanfuda)* and then exchanging them with fellow enthusiasts at regular *nōsatsu* club meetings (n*ōsatsu-kai*). After the Meiji Restoration and the move towards modernization and Westernization, the practice of meeting to exchange votive slips largely died down. However, sparked by a sense of nostalgia for Edo culture and traditions, *nōsatsu* exchange clubs became active again and even spread outside of Tokyo to other urban areas, such as Osaka and Kyoto.

The University of Oregon's Knight Library holds one of the only collections of *nōsatsu* outside of Japan and is now in the process of digitizing the entire collection of over 100 albums (approximately 4,000 individual images) adding metadata for each image and uploading the records to the Oregon Digital Archive (https://oregondigital.org/sets/gb-warner-nosatsu). Because most of the text is written in early modern Japanese style script, (*kuzushiji*) the process of inputting accurate and robust metadata has been relatively slow and at this point only two of the albums have been digitized, cataloged and uploaded. To speed up the pace of progress of this project, the Japanese Studies Librarian at the University of Oregon has started working in collaboration with the Digital Scholarship Center to identify crowdsourcing transcription software that will allow collaboration with specialists in early modern Japanese culture as well as *nōsatsu* connoisseurs to add and edit metadata records for the UO's nōsatsu collection.

The two *nōsatsu* enthusiasts who collected most of the votive slips in the UO's collection came from notably different backgrounds: Frederick Starr, an anthropology professor at the University of Chicago, who became intrigued with *nōsatsu* during his first visit to Japan in 1904 and Sato Masao, a self-employed woodblock printer, who was active in the *nōsatsu-kai* world from ca. 1920 to 1994.

Starr first joined the *nōsatsu-kai* milieu around 1910 and then became so passionate about attending meetings and collecting and exchanging votive slips that he became known as o-fuda hakushi (Professor Ofuda).[1] His votive slip collection included not only the *nōsatsu* he received from other collectors at regular meetings, but also several albums dating back to the late Edo period, with prints by prominent artists such as Utagawa Hiroshige and Utagawa Yoshitsuna.

After his death in 1931, Starr's *nōsatsu* collection was inherited by his sister, Lucy, who lived in Seattle, Washington. Since, she, apparently, had no interest in keeping the collection she contacted Gertrude Bass Warner, the primary benefactor of the University of Oregon's Art Museum and a woman who was well known for her travels in East Asia and zeal for collecting Asian art and artifacts. After negotiating via mail, Warner and Starr agreed upon a price for the *nōsatsu* collection and Starr's votive slips moved from Seattle to the University of Oregon in the early 1940s.

Sato, in contrast to Starr, who was a Western scholar [2] and complete outsider to Japanese culture, was, by the very nature of his profession as a printer and in his enthusiasm for votive slips, a sort of classic example of a typical *nōsatsu* club member. Since Sato's engagement with *nōsatsu* spanned several decades, the addition of his collection to the Starr materials gives the University of Oregon a fairly comprehensive chronological coverage of the history of *nōsatsu*, ranging from the late Edo period (ca. 1840s) to the Heisei period (ca. 1994) with images

depicting a wide variety of topics and themes, many of which harken back to traditional Edo culture: Scenes from famous literary works; Kabuki; Famous places in Edo; Folklore; Street performers; Sumo; Religious sites and pilgrimages. At the same time, *nōsatsu* images from, especially, the Meiji to early Showa era also provide glimpses of the massive changes that occurred in Japan after the Meiji restoration as well as reportage on current events. One series shows a, sort of, short history of the Westernization and modernization of Japan, with prints depicting telephones, electric lights, trains, subway stations and other newly established features of contemporary life. Another series focuses on the aftermath of the Great Kanto Earthquake of 1923 as citizens tried to recover from the damage and chaos caused by the earthquake and the massive fires that accompanied it.

As the digitization, cataloging and crowdsourcing of the UO's rare and culturally valuable collection proceeds, the primary objective is to form a network of Japanese Studies scholars, museum curators, academic librarians and *nōsatsu* specialists that transcends borders by making the images and metadata widely and freely available and enabling researchers to add and enrich metadata. On a local level at the University of Oregon, a *nōsatsu* research group, which meets on a regular basis, has been organized. The members, including a professor of early modern Japanese literature, a Japanese art history professor, two graduate students in Japanese studies, and the Library's Japanese Studies Librarian and Japanese Cataloger, meet to examine individual nosatsu albums, translate the cursive script and record metadata for cataloging. This group participated in a two-panel session devoted solely to *nōsatsu* at the Association for Asian Studies on the Pacific Coast Conference in 2017, which provided the opportunity to meet with other scholars interested in *nōsatsu* and opened up new channels for further expanding the network of *nōsatsu* experts. Finally, the Library received funding to hire a graduate student for the 2017-2018 academic year to work 20 hours a week on the University of Oregon's *nōsatsu* project. The student, working under the joint supervision of a Digital Scholarship Center staff member and the Japanese Studies Librarian, will be charged with identifying the most appropriate content management system for the project, recruiting scholars and *nōsatsu* aficionados to participate in the metadata crowdsourcing effort and training participants in the use of the content management system.

In addition, this collection is rife with the potential for future digital humanities projects, such as charting networks of Edo, Meiji and Taisho period collectors by the groups they belonged to or mapping Starr's extensive travels around Japan as he made pilgrimages, toured the old Tokaido Highway and barnstormed across Japan.[3]

Language Documentation and Standards in Digital Humanities: TEI and the documentation of Mixtepec-Mixtec

Jack Bowers (Austrian Academy of Sciences), Laurent Romary (INRIA)

This paper discusses an ongoing language documentation project covering the Mixtepec-Mixtec variety of Mixtec (iso 639-3: mix)[1]1 among the primary facets of which are to: create an open source body of reusable and extensible multimedia language resources (LR) encoded in TEI and in accordance with the recommendations of Bird & Simmons (2003); use the above in order to further the knowledge of all aspects of the language itself within the fields of linguistics and lexicography by producing empirical corpus-based descriptions and analyses of various aspects of the language's features; demonstrate and test the application and utility of descriptive features from cognitive linguistics theory[2] in the annotation of the LR; demonstrate and evaluate the application of encoding and description standards on a collection of lexical and knowledge resources for an under-resourced non-indo-european language.

Thus far, the majority of the project's data are collected from recordings of consultation sessions with native speakers, and a collection of children's texts in the language's working orthography published by the Summer Institute of Linguistics (SIL) Mexico. The recorded speech has been transcribed using Praat (Boersma and Weenik, 2017), then converted to TEI-XML using XSLT. The relatively small dataset [3]3 allows for ambitious goals with regards to the feasibility of applying an extensive degree of different annotation/markup to our dataset.

Concretely, the output include a collection of primary and secondary resources, specifically: audio files, digital TEI dictionary with English and Spanish translations, etymological information (where possible), TEI digital editions of the SIL Mexico publications with multi-level annotations, semantically and grammatically annotated corpus of all available original and external written LR, TEI utterance files with transcriptions and annotations, evidence based inventories and description of lexical features using TEI feature structures. Where possible, we make use of open source standard vocabularies such as ISOcat[4], which is the local reference point for the corpus annotation, feature structure inventories are created for the following lexical features: phonetics/phonology, morphophonology and morphosyntax and morphosemantics, a conceptual inventory for semantic profiles of Mixtepec-Mixtec vocabulary to be based in Linked Open Data resources such as dbpedia, Wikidata, etc.

As is typical in language documentation and lexicography, the organization and representation of the project's data involves collecting and integrating a wide array of linguistic, geographical, cultural, historical, and scientific information. While all of these things have been addressed before, there is less precedent for them all being attempted together within the TEI framework. Whereas the TEI has been widely applied to projects involving mostly indo-european languages, there has been few attempts to apply it to under described and indigenous languages; Czaykowska-Higgins et al. (2014) is the most noteworthy to date (see also: Bates & Londsdale 2009; Thieberger 2013). There has also been a lack of applications of the TEI to tonal languages such as Mixtec.

---

[1] Mixtepec-Mixtec ('Sa'an Savi' : 'rain language') is an Otomonguean spoken by roughly 9000-10000 people in the Juxtlahuaca district of Oaxaca, and parts of the Guerrero and Puebla states of Mexico. Most of the spoken data collected in this project originate from consultation sessions with two native speakers from a town called Yucanani (17.30083, -97.89389), which is part of the municipality of San Juan de Mixtepec with a population of around 500 people.

[2] In particular concepts from Langacker 1987, among others will be used and cited in full in final paper

[3] Thus far the project contains roughly 1000 original and collected open source sound files of various length, the vocabulary collected from which an combination with the SIL documents is expected to be in the low thousands.

[4] The ISOcat (ISO TC37) is currently undergoing a process of remodeling, editing and migration to a new management system 'TermWeb' (Wartburton, 2015).

Additionally, applying these methods, technology and standards to MIX is also challenging due to a lack of established linguistic description[5] and external language data; the orthography is not fully conventionalized[6], and it does not represent the lexical tone. This results in an extremely large number of homographs. The combination of these issues and the significant amount of phonetic variation encoded in the data makes glossing from sources, and applying NLP processes to search, retrieve and analyze highly difficult.

As mentioned, a major goal of this work is to further the body of knowledge of all aspects of the language's features. Given that linguistic subfields do not exist separate from one another, and they interact and overlap both synchronically and diachronically, it is important that the markup and annotation strategy used is able to capture all of these phenomena. Evidence for this can be seen in the fact that in MIX, changes in tone can indicate such morpho-semantic features as: negation, tense, person, possession or accompaniment. Thus we need a markup system that can handle overlapping features flexibly.

For both text sources and transcribed utterance files, standoff annotation provides a flexible means in which to flexibly and efficiently express the necessary linguistic and conceptual features in our data, including linguistic interface phenomena. For this purpose we adopt the recommendations of ISO 24624 (2016) and Schmidt (2011) as a baseline for our standoff annotation of spoken language transcriptions, integrating the ongoing work of Bański et al. (2016) regarding an expansion and refinement of the TEI standoff annotation system.

The need to process our spoken language data from sound files while providing an evidence based description of the language's phonetic and phonological systems requires the use of software tools that enables us to create annotations, and produce data that can be transferred to TEI. However, given our specific needs, there is still no perfect choice. Due to a combination of legacy reasons, and the desire to provide a quantitative body of evidence for the description of MIX phonology, we use Praat, despite the fact that this software is primarily geared towards the fields of applied phonetics/phonology, and is rarely used in digital humanities. Accordingly it lacks options for TEI or even basic XML output, and doesn't have the capacity to include key metadata with transcriptions and sound files. In our paper we will elaborate on why we chose Praat over other tools such as EXMARaLDA (Schmidt, 2004).

In our paper we describe basic facets about the MIX language, the data and give examples of markup methodologies for each stage of the workflow. Along the way we point out issues encountered with the relevant markup standards, vocabularies, resources and tools, discuss future and/or ongoing efforts to find solutions. Additionally, we provide insights into how we have attempted to navigate the need to converge the technical needs of the data, with the need to seamlessly describe the language's features according to relevant theoretical linguistic principles. Where the TEI and other relevant standards do not meet our needs, we provide concrete use cases and input for proposed improvements. Our efforts attempt to not only provide a lasting and reusable set of resources for the MIX language, but also to make strides towards bridging the gap between lexicography, language documentation, theoretical linguistics, computational linguistics and digital humanities.

Works Cited

[1] Bański, P. (2010). Why TEI stand-off annotation doesn't quite work. In *Balisage: The markup conference*.

[5] There have been previous descriptions of MIX phonology (see: Pike & Ibach 1978; Pastor 2005, 2010; Pastor & Beam de Azcona 2004a, 2004b); however none of these provided any corpus, acoustic or quantitative data. Our analysis has some significant differences that we will articulate in future publications based on empirical evidence.

[6] The executive council of the Academy of the Mixtec language has established a working blueprint for a standardized orthography for Mixtec languages (Consejo Directivo de Ve'e Tu'un savi. A. C, 2011, n.d.). However given the diversity in the varieties of Mixtec, some of which are not mutually intelligible, this directive is far from establishing a systematic governance of orthographic standards. The orthography currently used in the project is based on the de-facto conventions used by SIL Mexico.

[2] Bański, P., Gaiffe, B., Lopez, P., Meoni, S., Author, L., Schmidt, T., Stadler P. and Witt, A. (2016). Wake up, standOff! Presented at the TEI Conference and Members' Meeting, Vienna. https://hal.inria.fr/hal-01374102
[3] Bański, P., & Przepiórkowski, A. (2009). Stand-off TEI annotation: the case of the National Corpus of Polish. In *Proceedings of the third linguistic annotation workshop* (pp. 64–67). Association for Computational Linguistics.
[4] Bates, D., & Lonsdale, D. (2010). Recovering and updating legacy dictionary data. In *Proceedings of the 44th Annual International Conference on Salish and Neighboring Languages (ICSNL)* (Vol. 27, pp. 1–12).
[5] Boersma, P., & Weenik, D. (2017). Praat, a system for doing phonetics by computer (Version 6.0.28).
[6] Consejo Directivo de Ve'e Tu'un savi. A. C. (2011). Pronunciamiento de Ve'e Tu'un Savi, A. C. "Academia de la Lengua Mixteca" respecto a Tu'Un Savi (Lengua de la Lluvia). Retrieved from https://goo.gl/mnLrWt
[6] Consejo Directivo de Ve'e Tu'un savi. A. C. (n.d.). Ndusu. INALI. Retrieved from http://www.inali.gob.mx/pdf/ndusu.pdf
[7] Czaykowska-Higgins, E., Holmes, M. D., & Kell, S. M. (2014). Using TEI for an Endangered Language Lexical Resource: The Nxaʔamxcín Database-Dictionary Project. *Language Documentation and Conservation*, 8, 1–37.
[8] Drude, S. (2012). Digital Grammars -- *Integrating the Wiki/CMS approach with Language Archiving Technology and TEI*. University of Hawai'i Press.
[9] ISO-24624. (2016). Language resource management — Transcription of spoken language.
[10] Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites* (Vol. 1). Stanford university press.
[11] Paster, M. (2005). Tone Rules in Yucanani Mixtepec Mixtec (p. 13). Presented at the SSILA meeting, Oakland, Ca: SSILA.
[12] Paster, M. (2010). The role of homophony avoidance in morphology: A case study from Mixtec. In *13th Annual Workshop on American Indian Languages* (Vol. 21, pp. 29–39).
[13] Paster, M., & Beam de Azcona, R. (2004a). A Phonological Sketch of the Yucunany Dialect of Mixtepec Mixtec. In *the 7the Annual Workshop on American Indian Languages* (pp. 61–76). UC Santa Barbara.
[14] Paster, M., & Beam de Azcona, R. (2004b). Aspects of tone in Yucunany Mixtepec Mixtec. Presented at the Conference on Otomanguean and Oaxacan Languages, University of California, Berkeley.
[15] Pike, E. V., & Ibach, T. (1978). The phonology of the Mixtepec dialect of Mixtec. *In Linguistic and Literary Studies in Honor of Archibald A. Hill* (Descriptive Linguistics, Vol. 2, pp. 271–285). The Hague: Mouton.
[16] Schmidt, T. (2004). Transcribing and annotating spoken language with EXMARaLDA. In *Proceedings of the LREC-Workshop on XML based richly annotated corpora, Lisbon 2004*.
[17] Schmidt, T. (2011). A TEI-based Approach to Standardising Spoken Language Transcription. *Journal of the Text Encoding Initiative*, (Issue 1).
[18] Thieberger, N. (2013). Reusing manuscript vocabularies, an example from Western Australia. Presented at the Conference presentation at the 3rd International Conference on Language Documentation and Conservation (ICLDC 3), University of Hawai'i.
[19] Wartburton, K. (2015). Re-modelling the ISO T37 data category registry: current status, challenges, and theoretical reflections. *Deutscher Terminologie Tag - DTT*, (1).

Application of the Concept of “Linkbase” for the Digitalization of Linguistic Resources and Analysis

Yona Takahashi (University of Tsukuba)
Masakatsu Nagai (University of Tokyo)
Toshihito Waki (University of Tsukuba)

The presenters’ goal was to establish a new linguistic field and its methodology based on a photographic image database for written materials. Specifically, two types of textual data were created and stored into a database. Since 2015, the authors have begun to develop a bilingual database of the Egyptian-Akkadian treaty texts written in hieroglyphics and cuneiform, which were made by the Egyptian empire with the Hittite empire in the fifteenth century B.C.E., to analyze the grammatical correspondences between the Egyptian and Akkadian languages. In addition, since 2016, the authors have started to develop another digital archive of the literary text, “Epic of Gilgamesh,” which was written on cuneiform tablets and is widely known for more than two thousand years in the Ancient Near East. Since there are various editions of the epic which remained from each era and each region, it is necessary to compile them to describe their relationship.

The primary issue which needs to be considered is: how to describe many-to-many correspondences between resources as digital data? In the bilingual database case, the grammatical correspondences are not usually one-to-one, but many-to-many due to the differences in the grammatical systems between the two languages in order to clarify the linguistic features by contrastive analysis. In the case of digital archiving for the literary text, it is often that a part in a previous edition was divided into some divisions and rearranged in the subsequent editions, and vice versa. In order to comprehend continuous editing done by the ancient poets, the authors attempted to visualize the intermittent and many-to-many correspondences among the editions or their internal structures. For these reasons, a standard framework was necessary to describe the many-to-many correspondences.

To resolve this issue, the authors adopted the concept of “linkbase,” originally proposed in XLink 1.0. A linkbase is a "link database" which is a collection of third-party links where neither the starting resource nor the ending resource is local. Utilizing this linkbase, the authors would be able to express links which reside in a location separate from the linked resources. Linkbase makes link management easier by concentrating on document management while separating a number of link collections from the related documents.

However, the authors did not adopt the whole specification of XLink. Furthermore, the concept of “@actuate” (for the representation of the link) not need. Even XML syntax is not necessarily required. The main point is that a “link database” can be treated as the same object as a “document database” and an “image database.” Therefore, written materials, photo images, and linking relations can be digitalized uniformly, with accustomed digital tools such as Microsoft Excel, OpenOffice Calc, etc. at a low learning cost.

Then, the authors created a small web-based collation tool using HTML and XSLT to visualize the many-to-many links, in order to verify the complex linking structures and, if necessary, correct the data. The data was then provided in a highly sharable XML format conform to Text Encoding Initiative (TEI), not only for the document but also for the linkbase. For this purpose, TEI has <linkGrp/> element (§ 16.1) and Pointing mechanisms (§16.2). Especially the latter mechanisms allowed us to utilize any types of linking representation by converting them into canonical URI/IRI references (§ 16.2.5).

The semantics of the links (i.e. “@arcrole” in XLink) is still an open issue. Only two types of link are required for our purpose: direct correspondence and indirect/informative correspondence. As a consequence, it is sufficient for us to prepare two independent linkbases so far, by set

aside a complicated typology of links. A further discussion might be needed for the comprehensive digitalization in our databases.

The concept of “linkbase” does not depend on the methods or operations to utilize data. It allows us to choose a suitable data format for our own purposes, and leads us to achieve interoperability of the various open data by linking, even by linking of “link databases.” Further discussions and use cases on linkbase will be useful for our future.

TEI/XML Markup of Engi-shiki as Research Platform for Historians of Ancient Japan

**Naoki Kokaze (University of Tokyo),
Kiyonori Nagasaki (International Institute for Digital Humanities),
Makoto Goto, Yuta Hashimoto (National Museum of Japanese History),
Masahiro Shimoda, Albert Charles Muller (University of Tokyo)**

As part of the collaborative research project with the National Museum of Japanese History, this paper aims to report a joint research achievement through examples of TEI/XML markup, based on 'Transactionography,' for an ancient Japanese historical sources called *Engi-shiki*, which was compiled as an administrative manual in the tenth century.

According to one of the most influential historians on the *Engi-shiki*, Toshiya Torao, detailed rules on society and administration during that period were regulated very precisely. For instance, there were detailed regulations covering rituals and festivals held in various parts of Japan, designation of offerings of rituals, tributes and taxes paid to the *Ritsuryō* government, and allocation of Shōzei and Kugai (rice plants used as funds which were distributed to each administrative region). Due to the nature of the descriptions of such diverse details, the research subjects tend to be subdivided, and thus it is difficult to obtain comprehensive views on society at the time.[1]

In order to offer a possible solution for this historiographical obstacle, this presentation will provide some visualizations, such as a geographical comparison of each administrative region based on its financial scale, and variations in the processing methods of major products, e.g. abalone, through the text markup using the TEI/XML methodology 'Transactionography,' first developed by Drs. Tomasek and Bauman.[2]

This is an extended model of TEI used to mark up appropriately various kinds of Historical Financial Records (HFRs) containing histories of the exchange of goods and services, such as receipts or account ledgers.

In this paper, two examples can be addressed as a result of text markup of *Engi-shiki*. At first; when it comes to marking up Shōzei and Kugai rice plants, they can be expressed as budgets that each region could utilize for its administration.

Markup. 1: Shōzei and Kugai Rice Plants

```
陸奥国、<measure xml:id="mut1" commodity="正税" quantity="630000"
unit="束">正税六十万三千束</measure>、<measure xml:id="mut2"
commodity="公廨" quantity="803715" unit="束">公廨八十万三千七百十五
束</measure>
```

Taking into consideration the concept of 'Transactionography,' we define the meanings of Shōzei and Kugai budgets as transfer of measure from people to each county. In this markup strategy, people can be understood as inhabitants of the counties.

[1] Toshiya Torao, Engi-shiki no Kenkyū (A Study on Engi-shiki), Tokyo: Yoshikawa Kōbunkan, 1995.

[2] Kathryn Tomasek and Syd Bauman, 'Encoding Financial Records for Historical Research', Journal of the Text Encoding Initiative [Online], Issue 6 | December 2013, Online since 22 January 2014, connection on 1st May 2017. URL : http://jtei.revues.org/895 ; DOI : 10.4000/jtei.895.

```
<hfrs:listTransaction>
      <hfrs:transaction><!- Shōzei distributed to Mutsu county->
            <hfrs:transfer fra="#people" til="#陸奥国">
                  <measure sameAs="#mut1"/></hfrs:transfer>
      </hfrs:transaction>
      <hfrs:transaction><!- Kugai distributed to Mutsu county->
            <hfrs:transfer fra="#people" til="#陸奥国">
                  <measure sameAs="#mut2"/>
            </hfrs:transfer>
      </hfrs:transaction>
</hfrs:listTransaction>
```

To make use of these inline and standoff markup texts, we connect them with geographical information of each county seat, thus obtaining a visualization such as follows:

Figure. 1: Financial Comparison of Administrative Regions based on the Amount of Shōzei and Kugai Rice Plants[3]

Second, we mark up the products. In the *Engi-shik*i, we can see a variety of products emerging in various parts of detailed regulations used as offerings or taxes. Among them, abalone is one of the most circulated foods.

Markup. 2: Hierarchical Marking up of Abalone

```
<measure type="abalone" xml:id="鰒 1" commodity="鰒" quantity="5"
unit="両">鰒</measure>・堅魚各五両
```

One important point in this markup is multilingualism. To offer the variation of products name (which means the variety of processing method of food as well as the places and regions of production), all terms referring to abalone should be marked up with <measure> element containing the type ="abalone" attribute written in English, and subsequently, the commodity attribute can be used to address precisely the name of products in Japanese. Consequently, we can comprehend the variety of names signifying abalone, as follows:

[3] This mapping visualization is created with cartoDB (https://carto.com/).

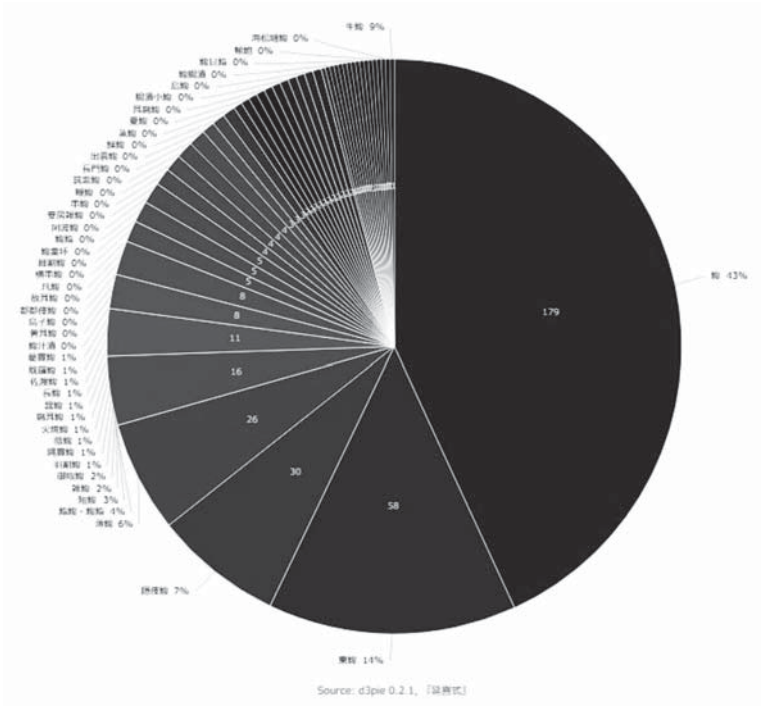


Figure. 2: Variations of the name referring to Abalone in the *Engi-shiki*

Apart from these example data manipulations, the important contribution of our study will be to formulate a research platform enabling historians to offer various aspects of society and administration in ancient Japan, according to the concerns of individual researchers or research projects.

As a future vision of collaborative projects with the National Museum of Japanese History, it is expected to clarify the co-relationships between images and actual texts by exposing such a text database together with the image database based on IIIF (International Image Interoperability Framework), and to expand the possibility of making the best use of historical materials. The progress of our ongoing collaborative research project should make it possible to disseminate the accumulation of Japanese ancient historical research more effectively and internationally.

Ontological Engineering in Philosophy. Example of Phenomenological Data

Raphael Kur (Jagiellonian University)

Digitization of philosophy is already a fact. Millions of texts have been archived and made available to users, typically on the world wide web. In my paper, I would like to deepen the meaning of the term “digitization of philosophy”, since the process of digitization is not limited to text scanning, resulting in a graphic file (bitmap). Scanning is followed by a series of important processes like data (graphic) compression and an automatic process of recognition of characters and words using the OCR (Optical Character Recognition) method. Still, the most important element in the whole process of digitization is the final product made of parts that underwent selection, organization, and categorization. Just like any genre of literature requires navigation, so does philosophical literature, to make it structured and user-friendly.

On one hand, the idea I am presenting in this article is not novel. Some institutions are already working on this topic, using ontological (philosophical) ideas and computer methods, usually applied to categorize knowledge in biomedicine. This particular type of research is referred to as ontological engineering. Institutions active in this area are for example Metaphysics Lab at Stanford University, Ontology Research Group at University of Buffalo (New York), or Laboratories for Applied Ontology in Italy (Trento). On the other, it is only the Indiana Philosophy Ontology (InPhO) project, carried out at the Indiana University Bloomington that applied ontological categorization covering philosophy in its entirety, albeit in a very general manner, operating basic information.

Ontologies are representations of entities and relationships between them. Engineering means a technical proposal for solving a concept. Many of these ontologies are based on natural language though the philosophical text corpus is highly specific, often containing ambiguous terms. Therefore, building a simple, yet transparent structure, and at the same time reflecting the richness and meaning of philosophical concepts is not a simple task. Since ontological engineering of philosophy is not that widespread, we should expect several approaches and concepts in this area.

It would seem that the best place to start categorizations is from the most characteristic elements of a genre of literature, in this regard from one philosophical concept and its evolution, including relationship with other philosophical fields. Accordingly, what I would like to present is a proposal of a selected philosophical trend on a general, meta-level.

Contemporary philosophical literature consists of a large number of diverse publications, ranging from dictionaries, encyclopedias, books, to articles and abstracts. All these contain keywords and indexes, required when searching for information within the text, however, a much higher level of order is required, general and abstract enough to be able to connect even the most unassociated ideas. This is especially important today when we are flooded with information, likewise in the field of philosophy. Ontological organization in the digitized world of philosophy is needed today more than ever.

Philosophy has many branches of ideas, and various kinds of divisions of the general sub-disciplines (metaphysics, ethics, aesthetics, axiology, political philosophy, philosophy of science, philosophy of mind, etc.) derived from authors of (e.g. Platonism, Confucianism, Schopenhauerian philosophy, Kantianism; phenomenology from the point of view of Husserl, Heidegger, Ingarden, etc.). Combinatorics in this regard is very diverse as a result of many variables and aspects (e.g. geographical, cultural, ethnic, religious, linguistic, etc.), not forgetting that the selected concepts may be found by using a specific time and location (e.g. the nineteenth-century aesthetics and neuroaesthetics, Japanese and Polish phenomenology, etc.). It makes sense, therefore, to present this type of information in the form of a tree. This method

has long been used in logic and works in computer science. This type of structure of information is accessible and easier to formalize at different levels (from the most general to detailed), followed by implementation of programming. Consequently, in the near future, when we attain large collections of philosophical information, programs, operating efficiently on large bodies of text will be more useful in specific, specialized research.

These types of hierarchical structures should be separated by specific ontological concepts and include specific links/relations. The following diagram is a proposed example. Phenomenology as a philosophical movement, categorized by authors, varieties of thought (ideas), geographical extents, and temporal frames. By creating general schemes, we can create additional detailed levels (branches), then try to implement their empirical methods using logical indicators, in this way we decrease the proximity from methods of programming.

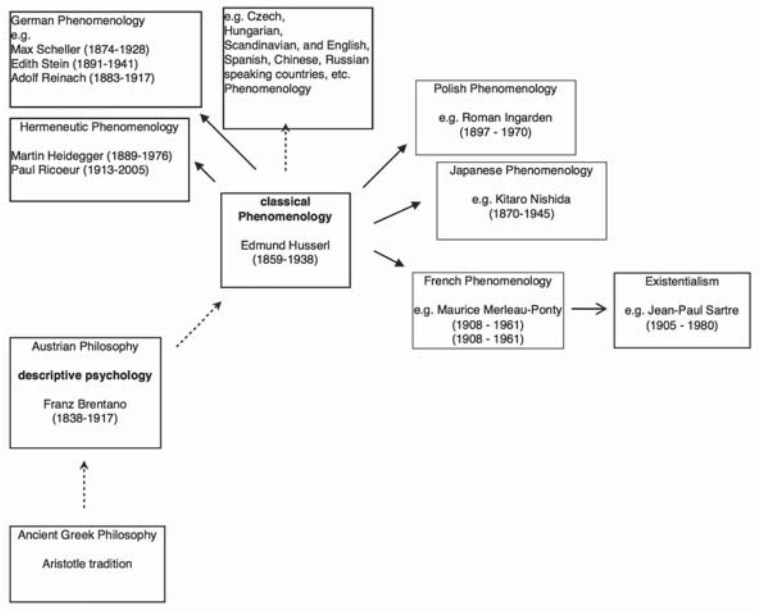


Figure. 1: Phenomenology segmented along geographical, philosophers and temporal axes with the resultant cross-segmentation

References

[1] Arp, R., Smith, B., & Spear A.D., (2015). *Building Ontologies with Basic Formal Ontology.* Cambridge, Massachusetts, London; MIT Press.

[2] Smith, B., (2003). Preprint version of chapter “Ontology”, in L. Floridi (ed.), *Blackwell Guide to the Philosophy of Computing and Information.* Oxford; Blackwell, 155–166.

[3] Sowa, J. F., (2000). *Knowledge Representation. Logical, Philosophical and Computational Foundations*. Pacific Grove, CA; Brooks Cole Publishing Co.

[4] Russell, S., Norwig, P., (2010). *Artificial Intelligence: A Modern Approach*. Prentice Hall.

Development of a title authority database for Chinese classic works

Maiko Kimura (University of Tokyo)

Background

The *Functional Requirements for Bibliographic Records* (FRBR) and its revised version, the *IFLA Library Reference Model* (LRM)[1] demonstrate a new model for library cataloging. However, it is questionable whether the FRBR model is applicable to Chinese classics in libraries, as the model was developed by The International Federation of Library Associations and Institutions (IFLA), whose member libraries are mostly from Western countries[2]. As such, feasibility research on the adaptation of the FRBR for Chinese classics is lacking.

Although the LRM emphasizes authority control, authority data for work titles has hardly been created in Japan, as the *Nippon Cataloging Rules* (typical Japanese cataloging rules that have not yet adopted the FRBR) do not require it. Currently, no comprehensive data is available in Japan for Chinese classic works.

Purpose

The purpose of this study is to create comprehensive title authority data for Chinese classic works, adopting the simplified FRBR model for Chinese classics. In addition, the author defines metadata vocabularies to release the data as Linked Open Data (LOD).

Model Design

As shown in Figure 1, compared to the relationships among entities representing products of intellectual or artistic endeavor in the FRBR model, in the proposed model, which is specifically for Chinese classics, these relationships are simplified.

Although most bibliographic records created in library catalogs are at the Manifestation level (see Figure 1), Item-level records have been created for classic books, as the characteristics of each item such as interpolations, seals, binding, and so on are important in classic works. Therefore, libraries tend to create bibliographic records at the Item level for classic books, at least in Japan.

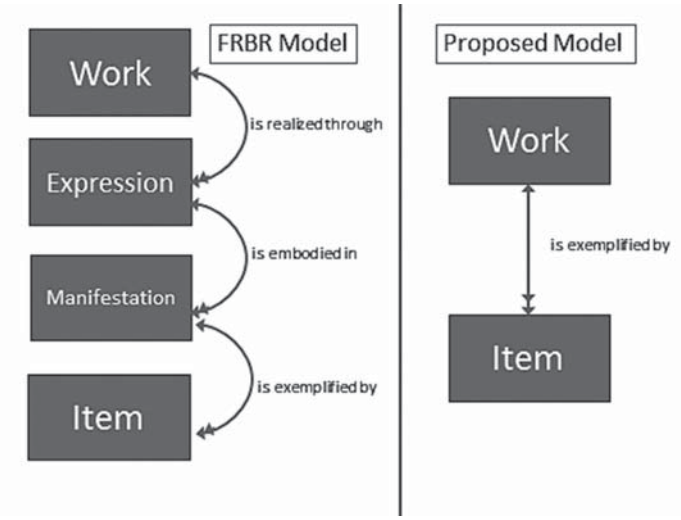


Figure 1: Comparison between the FRBR model and proposed model

In addition, the concepts of Expression and Manifestation are not always appropriate for Chinese classics. For example, although a partial re-carving of a printing block is important for users of Chinese classics, it is unclear whether we should perceive only one spot of the re-carving as a difference in Manifestation in the FRBR.
Bibliographers and researchers organize published, printed, or revised Chinese classic works in traditional classifications or orders. Library communities should carefully consider if we can simply change the results of their research into the series of Work, Expression, Manifestation, and Item. The proposed model only adopts Work and Item, leaving sufficient space for the utilization of research results by bibliographers in the future.

Developments

Recording work titles

A title authority database for Chinese classic works, namely the *KWMA-san*[3], was developed based on *Tong shu yi ming hui lu*[4], which lists various titles for the same work. Since the book does not show a representative title (preferred title) among several titles for a work, the author selected the representative title for each work according to the following order of priority. 1. The title that can return the most search results in the *Zenkoku Kanseki Database* is preferred.[5] 2. The title not used in other works is preferred. 3. The title that well represents the subject of the work is preferred. To distinguish different works sharing the same title, the second priority takes precedence over the first for some works. This means there is no duplication among representative titles. Both representative and non-representative titles are recorded in the database.
Tong shu yi ming hui lu is written in simplified Chinese characters. However, the author inputs all titles for each work in traditional Chinese characters, simplified Chinese characters, and *pinyin* (Romanization for the Chinese language) to improve data searchability.

Linking Work to Item

Here, CiNii Books[6] is searched using the representative title and non-representative titles for each work. Then, appropriate bibliographic records are selected and imported through the OpenSearch API of CiNii Books.[7] This means that each work has item titles that are exemplars of the work, allowing users to retrieve them using the title and alternative titles of the Item and Work titles. Although the main purpose of importing bibliographic records is to obtain the title and alternative titles of each item, author names, publication years, publishers, holdings, notes, and the URL of the bibliographic record in CiNii Books are also recorded in *KWMA-san*.
When CiNii Books does not return results, the link cannot be made, and only the work record exists in the database. As of June 12, there are 200 work authority records in *KWMA-san*, and 73 of these are without item titles.

Towards LOD

Figure 2 illustrates a current plan for metadata vocabularies to represent *KWMA-san*'s data. It adopts some vocabularies from the SKOS extension[8], DCMI Metadata Terms[9], BIBFRAME 2.0[10], and original defined vocabularies. The prefixes of the original vocabulary are <rep:> for representations and <kwma:> for the proposed model, and their URLs temporally do not exist. It differentiates transcription from the transliteration of each language, and considers a title in other languages as different titles. Thus, all titles other than a preferred title with a <skosxl:preflabel> property have a <skosxl:altlabel> property.

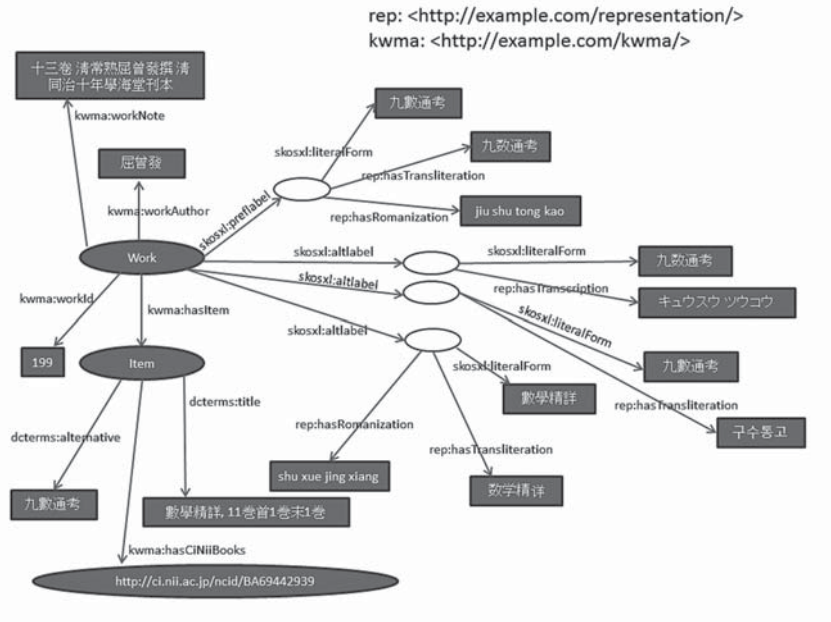


Figure 2: LOD vocabularies for *KWMA*-san

Acknowledgement

This work was partially supported by Konosuke Matsushita Memorial Foundation Research Grants 2016 and JSPS KAKENHI Grant Number JP 17J40023.

Reference

[1] IFLA. (2017). *IFLA Library Reference Model.* https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla_lrm_2017-03.pdf
[2] IFLA. (2015). *Annual Report 2014*. Retrieved from https://www.ifla.org/files/assets/hq/annual-reports/2014.pdf
[3] *KWMA-san*. Retrieved from https://zoshoin-db-zosan.herokuapp.com/works
[4] 杜信孚, 王剑编. (2000). 同书异名汇录. 江苏古籍出版社.
[5] *Zenkoku Kanseki Database*. Retrieved from http://kanji.zinbun.kyoto-u.ac.jp/kanseki
[6] National Institute of Informatics (NII). *CiNii Books*. Retrieved from http://ci.nii.ac.jp/books/?l=en
[7] NII. CiNii Books: Metadata and API:CiNii Books OpenSearch for Books & Journals. Retrieved from https://support.nii.ac.jp/en/cib/api/b_opensearch
[8] W3C. SKOS Simple Knowledge Organization System eXtension for Labels (SKOS-XL) Namespace Document:HTML Variant. (2009, November 27). Retrieved from https://www.w3.org/TR/skos-reference/skos-xl.html
[9] DCMI. DCMI Metadata Terms (2012, June 14). Retrieved from http://dublincore.org/documents/2012/06/14/dcmi-terms/
[10] The Library of Congress. BIBFRAME 2.0 Vocabulary List View. Retrieved from http://id.loc.gov/ontologies/bibframe.html

Universal Dependency for Modern Japanese

Mai Omura (National Institute for Japanese Language and Linguistics), Yuta Takahashi (Meiji University), Masayuki Asahara (National Institute for Japanese Language and Linguistics)

Universal Dependency (UD; McDonald et al., 2013) is an international project to develop cross-linguistically consistent treebank annotation for more than 60 languages. The resources by the project include ancient languages such as Ancient Greek, Coptic, Old Church Slavonic, and Sanskrit. The UD Japanese team are developing an annotated corpus of contemporary written Japanese (Tanaka et al., 2016). However, they have not developed the corpus for the ancient Japanese language. Constructing the UD compatible syntactic dependency annotation on modern Japanese enables to research linguistic typology with other languages diachronically.

We developed a dependency treebank for the 'Meiroku-Zasshi' corpus in the *Corpus of Historical Japanese* (National Institute for Japanese Language and Linguistics 2017). We chose six samples from the corpus because these samples are annotated for *Bunsetsu* (Japanese base phrase) boundaries, two morpheme units (Short Unit Word and Long Unit Word), and morphological information.

In the first step, we developed *Bunsetsu*-based dependency relations and coordinate structure annotation compatible with BCCWJ-DepPara standards (Asahara, 2016). One annotator, who is a graduate student in modern Japanese, annotated the all six samples. There are three phenomena in the modern period that the guidelines for contemporary written Japanese cannot cover. One phenomenon is inversion derived from classical Chinese (Figure 1). The second is the predicative adnominal adjective (Figure 2). In contemporary written Japanese, adnominal adjectives occur mostly in attributive usage. However, in modern written Japanese, both predicative and attributive usages are attested. The last one is concord between adverbs and postpositions (Figure 3). Note that, in the Japanese syntactic standard, the arc direction is from dependent to head.

Figure 1: Inversion

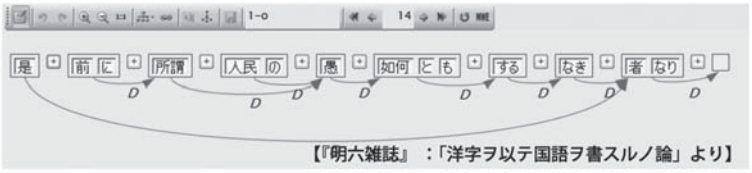


Figure 2: Predicative usage of adnominal "所謂"

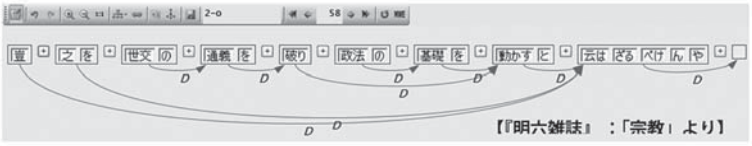


Figure 3: Adverb concord

In the second step, we annotated predicate-argument structures on the same samples.
Three annotators annotated the samples. The annotation standard is compatible with NAIST Text corpus (Iida et al., 2007) and BCCWJ-PAS. The target predicates (red background) are verbs, adjectives, and predicative nouns. We annotated subject (*-ga*, green background), direct object (*-o*, yellow background) and indirect object (*-ni*) for the predicates as in Figure 4. We also annotated the zero-pronoun information for the predicate when the subject was not overtly expressed. One issue in the modern language is functional expressions, which are not predicative expressions. The other issue is resolution of the zero pronoun. The annotator needs background information about the samples to resolve whether the subject is the first person, the second person, or others.

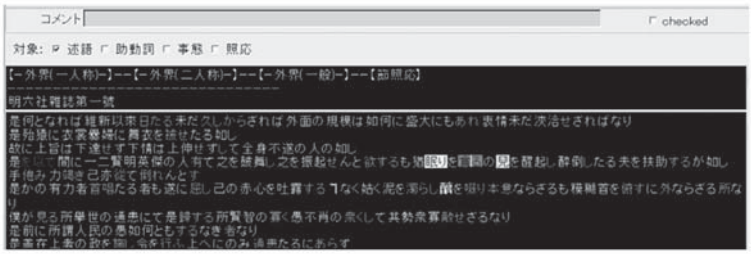


Figure 4: Predicate-argument structure annotation

In the third step, we converted these syntactic annotation standards into UD standard. The word unit of CHJ is based on the UniDic Short Unit Word. The unit is accepted as the UD standard for Japanese. We introduced conversion rules for the layers of morphology, dependency attachment, and dependency label. The morphology layer is based on a POS correspondence table between UniDic POS tagset of CHJ and UPOS (Universal POS in UD). The syntactic dependency attachments are determined by Bunsetsu-dependency relations as follows: (a) a head content word in a modifier Bunsetsu attaches to the head content word in the modifiee Bunsetsu; (b) non-head content words attach to the head content word in the same Bunsetsu; (c) a head functional word attaches to the head content word in the same Bunsetsu; and (d) non-head functional words attach to the head functional word in the same Bunsetsu. The head information of content or functional word is determined by the head rules in a dependency structure analyser CaboCha. Then, the dependency labels are based on conversion rules by the POS of the paired words, subtree, and predicate-argument relations. Figure 5 shows the converted dependency tree in Universal Dependency schema.
Note that the arc direction is from head to dependent in the Universal Dependency community.

Figure 5: An example of Universal Dependency tree for Modern Japanese

We still have some issues regarding coordinate structures which have low affinity to dependency relations, so, we manually modified such relations. In the poster presentation, we will present the overview of syntactic annotation procedures, guidelines, and data. In our future work, we will enter these syntactic annotations on 'Kokutei Tokuhon' and 'Taiyo Corpus'.

Acknowledgements

This work was supported by JSPS KAKENHI Grants Number 17H00917, 15K12888, and 17J03579 and a project of the Center for Corpus Development, NINJAL.

References

[1] R. McDonald, J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, K.Ganchev, K. Hall, S. Petrov, H. Zhang, O. Täckström, C. Bedini, N. B. Castelló, and J. Lee. (2013), “Universal Dependency Annotation for Multilingual Parsing”, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 92- 97.

[2] T. Tanaka, Y. Miyao, M. Asahara, S. Uematsu, H. Kanayama, S. Mori, and Y. Matsumoto. (2016), “Universal Dependency for Japanese”, Proceedings of LREC 2016, Tenth International Conference on Language Resources and Evaluation, pages 1651-1658.

[3] National Institute for Japanese Language and Linguistics. (2017), "Corpus of Historical Japanese" (Version 2017.3, Chunagon Version 2.2.2).

[4] M. Asahara, and Y. Matsumoto. (2016), “BCCWJ-DepPara: A Syntactic Annotation Treebank on the `Balanced Corpus of Contemporary Written Japanese’”. Proceedings of the 12th Workshop on Asian Language Resources, pages 49-58.

[5] R. Iida, M. Komachi, K. Inui and Y. Matsumoto. (2007), “Annotating a Japanese Text Corpus with Predicate-Argument and Coreference Relations”, Proceedings of the Linguistic Annotation Workshop, pages 132-139.

CBETA Research Platform: A Digital Research Environment for Studying Chinese Buddhist Literature in the New Era

Jen-jou Hung (Dharma Drum Institute of Liberal Arts)

In recent years, the emergence of digital humanities from traditional disciplines became a highlight in the field and received considerable attentions from scholars and specialists. Digital humanities researches focus on applying the application of current information technology to explore existing archival materials. The objective is to aid in the effort of humanities researchers to detect patterns hidden in the vast data set which conventional methodology fail to do, or for developing new approaches to understand the text.

Digitalization and textual analysis is one of the most sought of disciplines in the field of digital humanities. It is centered around the theme to digitalized ancient texts and presents them as searchable and computable electronic texts. Other than creating research platforms for reading purposes or databases for further exploration, through text mining applications based on these data, information between the lines might be revealed. This may contribute, especially to the study of ancient texts at various research levels.

In the field of Chinese ancient texts digitalization, the digitization of Buddhist scriptures has been regarded as a relatively complete and fruitful collection. The Chinese Buddhist Electronic Text Association (CBETA) has made the Chinese Electronic Tripi□aka Collection widely available for many years and provided a resourceful platform for the studies on Chinese Buddhist texts. In order to make these materials even more accessible and to provide advanced digital humanity tools, a new generation called the CBETA Research Platform (CBETA-RP) was developed by CBETA together with the Dharma Drum Institute of Liberal Arts.

CBETA-RP (http://cbeta-rp.dila.edu.tw/?lang=en) is the result of this project. It is designed to meet the research need of scholars with new tools. It is functionally tailored for three specific roles:

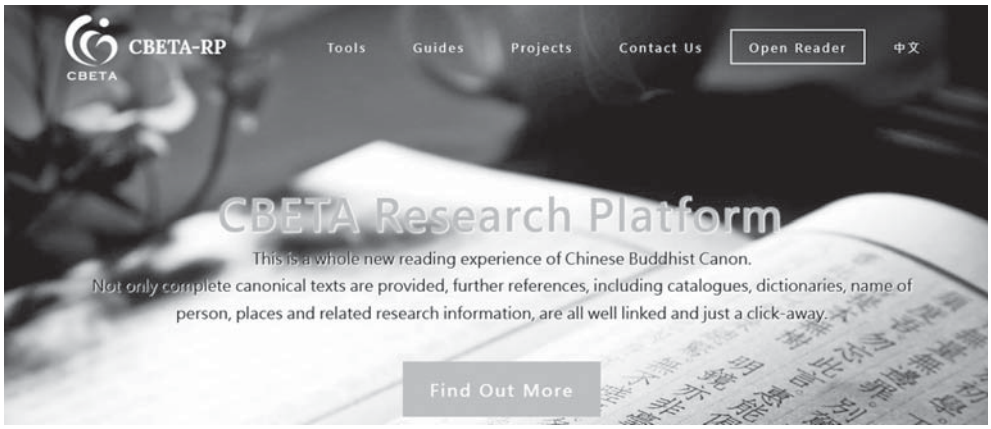


Figure 1: Homepage of CBETA Research Platform

(1) **Integrated Text Reader** (http://cbetaonline.dila.edu.tw/en/): It provides a whole new online reading platform with complete canonical texts and functional tools. Other than the entire CBETA database, functions such as Full-Text Search, Dictionaries Lookup, Term Usage Statistics, Integrated Bibliographic Database and Buddhist Studies Authority Databases are inclusive.

Figure 2: User interface of CBETA Online Reader.

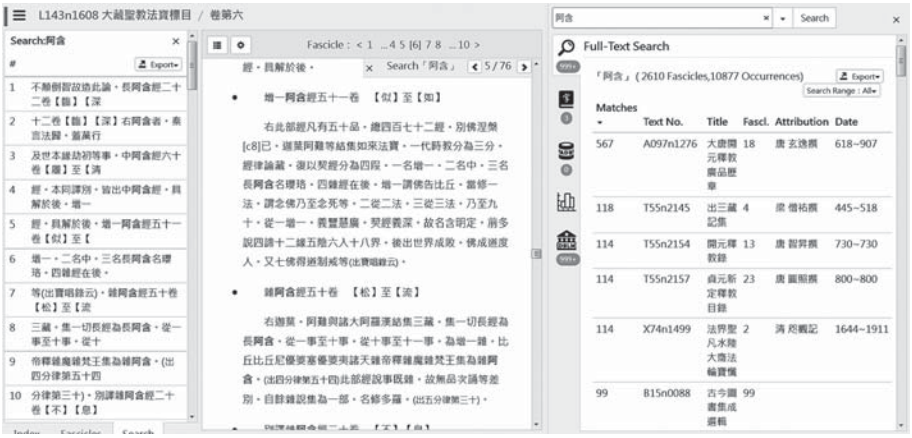


Figure 3: Full-Text Search results of CBETA Online Reader.

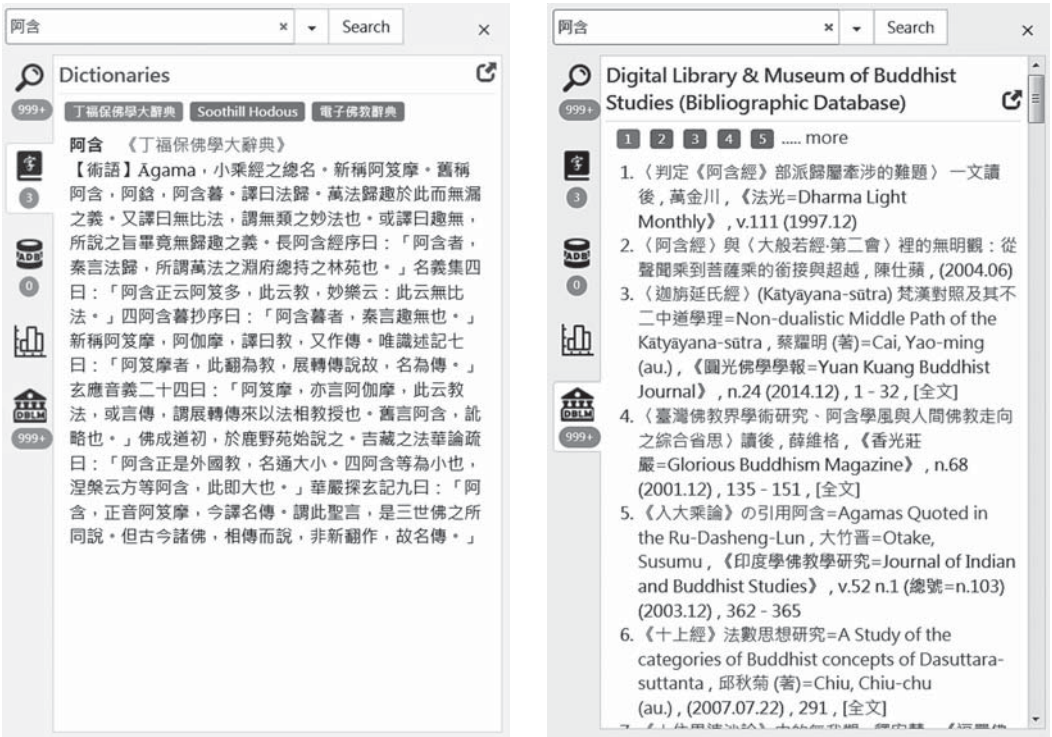


Figure 4: Term and bibliography search results of CBETA Online Reader.

(2) Search & Comparison (http://dev.dila.edu.tw/TaishoConcordance/?lang=en**):** Data comparison is the first step in data analysis. Through comparing a large number of digital data, researchers will be able to locate similarities and differences between data for in-depth studies. This project proposes to design various comparison tools for easy manipulation and constructive output, including information on corpus and the structure of scriptures etc.

		Dynasty	如是我聞	我聞如是	Remove?
1	☑	西漢 (-201 BCE ~ 9 CE)	0	0	
2	☑	新 (9 CE ~ 25 CE)	0	0	
3	☑	東漢 (25 CE ~ 220 CE)	5	0	
4	☑	三國 (220 CE ~ 280 CE)	6	1	
5	☑	西晉 (280 CE ~ 317 CE)	9	0	
6	☑	東晉 (317 CE ~ 420 CE)	461	225	
7	☑	南北朝 (420 CE ~ 589 CE)	1703	9	

Figure 5: Search and comparison interface

(3) Digital Textual Analysis Toolset: Text-mining is an important technique for text content exploration. With the aid of digital tools, important information on the textual structure and context could be uncovered. This project proposes to develop a series of text mining tools which would enable the user to fly higher. As this is an on-going effort, the result is yet to be presented.

In this Post, we will introduce the completed, the developing and the proposals of the CBETA Research Platform to researchers alike for constructive feedback.

Situational Effects on Functional Word Frequencies within Conversational Sentences in Japanese Novels

Hajime Murai (Future University Hakodate)

Introduction

In order to process story texts automatically utilizing artificial intelligence, it is necessary to identify relationships between linguistic characteristics and various attributes of those stories. There are many types of attributes that affect text style (e.g., genre, cultural and historical setting, social background, characters' personalities, scene mood…). If relationships between words in texts and key concepts can be identified, it may become possible to develop algorithms that interpret stories flexibly, just like a human being does. Moreover, those mechanisms could also be applied in automatic story generation systems.

Among the various elements of story texts, conversational sentences are a particular challenge for automatic processing. This is partly because colloquial language often include irregularities reflecting everyday usage of omissions and idiomatic expressions. Moreover, the way each character talks is often intentionally differentiated in general story texts [1]; this is a general technique used to help readers understand each character's personality. Generally speaking, readers of novel texts can expect the characteristics of each character's dialogue to convey information to be conveyed regarding various attributes (e.g., gender, age, temperament, social status) of the characters that appear in the work based on the characteristics of each character's dialogue. Therefore, even if there is no clue to the speaker's identity in descriptive text, most readers can accurately estimate who is speaking by reading only the conversational sentences.

Of the many attributes of conversational sentences in story texts, this paper focuses on their situational attributes, that is, attributes which indicate the characters' circumstances, like era, region, culture, and public or private context. These situational attributes are characteristic of a specific scene rather than belonging to individual characters (individually) across contexts. Because situational attributes don't directly contribute to character identification, their relationship with the characteristics of words appearing in the text has not been analyzed much in previous research for Japanese novels. However, situational attributes play a important role in creating a desired mood; therefore, their mechanisms should be clarified as well if we are to realize algorithms that can interpret stories flexibly.

Target Corpus

In order to analyze relationships between attributes and conversational sentences, a tagged-dialogue corpus of Japanese novels has been developed [2], based on random sampling of Japanese novel texts within the Balanced Corpus of Contemporary Written Japanese (BCCWJ) [3]. The Japanese novel texts, from among those classified as no. 913 (novels) in Nippon Decimal Classification, were extracted from the library-based (sub)corpus in BCCWJ; 100 texts were selected randomly. Conversational sentences were extracted from selected texts, and were manually tagged for speaker name, gendar of speaker, occupation of speaker, listener name, and relationship between speaker and listener. In all, 5632 utterances from the 100 Japanese novels were tagged. By combining utterances with the same speaker and listener, this was reduced to a data set of 887.

The situational attributes of era, area, and public or private situation were also tagged when found in story texts (Tables 1 and 2).

Table 1: Situational tags about area and era

Era / Area	Present	Past
Domestic	569	194
Foreign	28	82

Table 2: Situational tags about public or private situation

Family	90
Friend	82
Workplace	362
Other	353

Function Word Frequency and Usage Patterns

As characteristics of text style, the frequencies of different function words in utterances were adopted in this paper. Usage patterns of function words are differentiated in order to clarify speakers' personalities in these Japanese novels [1].

In the Japanese language, function words are mainly particles and auxiliary verbs. Therefore, in this paper, statistical significances for frequencies of particles and auxiliary verbs were analyzed, through the chi-squared test (p value is less than 0.01, Cramér's V is 0.10). Table 3 shows particles and auxiliary verbs with highly significant values (p value is less than 0.0001) on the chi-squared test for four pairs of area and era situational tags (present-domestic, present-foreign, past-domestic, and past-foreign). Because there were many words which showed significance, Table 3 describes frequently appearing words only. Table 4 does the same for the public–private division (p value is less than 0.01, Cramér's V is 0.14).

Table 3: Statistically significant words by area and era

	Significantly more	Significantly less
Present-domestic	*ta*, *desu*, *nai*, *yo*, *te*, *kara*, *teru*, *wa*	*ha*, *no* (case particle), *ni*, *wo*, *na*, *zu*
Present-foreign	*masu*, *desu*	*ta*, *desu*, *nai*, *yo*
Past-domestic	*ha*, *no* (case particle), *ni*, *wo*, *mo*, *na*, *zu*	*ne*, *teru*, *wa*
Past-foreign	*wo*	*masu*, *teru*

Table 4: Statistically significant words by public or private division

	Significantly more	Significantly less
Family	*te*, *yo*, *ne*, *kara*, *teru*, *no* (auxiliary particle)	*wo*, *desu*
Friend	*no*, *yo*, *na*, *teru*, *wa*, *keredo*	*ha*, *wo*, *masu*, *desu*, *zu*, *ga*
Workplace	*ha*, *no* (case particle), *wo*, *masu*, *ka*, *desu*, *ga*	*da*, *te*, *no* (auxiliary particle), *nai*, *yo*, *kara*, *teru*, *wa*, *no*, *keredo*

In Table 3, particles which adjunctively describe tones such as “Yo”, “Ne”, “Teru”, “Wa” are significantly more in Present-Domestic cell. Those particles are conversely significantly less in Past-Domestic cell. With respect to auxiliary verbs, words about politeness “Masu”, “Desu” are more in present division and are less in past division.

In Table 4, particles which connote pragmatic meaning, such as *te*, *yo*, *ne*, *teru*, *no*, and *wa*, are divided into three types by frequency pattern. *Yo*, *teru*, and *no* are all significantly high in both family and friend categories; *te*, *ne*, and *kara* are unique to the family division; and *na*, *wa*, and *keredo* are unique to the friend division. Meanwhile, auxiliary verbs conveying politeness are characteristic of the workplace division, implying that a professional way of speaking in Japanese includes polite phrases but excludes unnecessary connote pragmatic meanings. Moreover, particles that describe tones can be categorized based on the situations in which they appear.

Conclusions and Future Work

To analyze relationships between situational attributes of utterances and speech styles, this data set of randomly selected conversational sentences from a Japanese novel corpus was tagged and frequently appearing particles and auxiliary verbs were analyzed statistically. The results show clearly that era, area, and public–private differences affect language use in terms of politeness and tones. Further investigation with a larger corpus in future work could pull out more minute characteristics.

References

[1] Satoshi Kinsui, “Virtual Japanese: Mystery of Function Words”, Iwanami Shoten, 2003. (In Japanese)
[2] Hajime Murai, “Towards Agent Estimation System for Story Text Based on Agent Vocabulary Dictionary”, IPSJ Symposium series, Vol. 2016, No. 2, 2016. (In Japanese)
[3] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. “Balanced Corpus of Contemporary Written Japanese”. Language Resources and Evaluation, Vol. 48, No. 2, pp. 345-371, 2014.

Enabling digital humanities research and teaching through digital library APIs

Donald Sturgeon (Harvard University)

As digital libraries continue to grow in size and scope, their contents present ever increasing opportunities for use in data mining as well as digital humanities research and teaching. At the same time, the contents of the largest such libraries tend towards being dynamic rather than static collections of information, changing over time as new materials are added and existing materials augmented in various ways. Application Programming Interfaces (APIs) provide efficient mechanisms by which to access materials from digital libraries for data mining and digital humanities use, as well as by which to enable the distributed development of related tools. Here I present a working example of an API developed for the Chinese Text Project digital library (http://ctext.org) being used to facilitate digital humanities research and teaching, while also enabling distributed development of related tools without requiring centralized administration or coordination.

Firstly, for data-mining, digital humanities teaching and research use, the API facilitates direct access to textual data and metadata in machine-readable format. In the implementation described, the API itself consists of a set of documented HTTP endpoints returning structured data in JSON format. Textual objects are identified and requested by means of stable identifiers, which can be obtained programmatically through the API itself, as well as manually through the digital library's existing public user interface. To further facilitate use of the API by end users, native modules for several programming environments (currently including Python and JavaScript) are also provided, wrapping API calls in methods adapted to the specific environment. Though not required in order to make use of the API, these native modules greatly simplify the most common use cases, further abstract details of implementation, and make possible the creation of programs performing sophisticated operations on arbitrary textual objects using a few lines of easily understandable code. This has obvious applications in digital humanities teaching, where simple and efficient access to data in consistent formats is of considerable importance when covering complex subjects within a limited amount of classroom or lab time, and also facilitates research use in which the ability to rapidly experiment with different materials as well as prototype and reuse code with minimal effort is also of practical utility.

Secondly, along with the API itself, the provision of a plugin mechanism allowing the creation of user-definable extensions to the library's online user interface makes possible augmentation of core library functionality through the use of external tools in ways that are transparent and intuitive to end users while also not requiring centralized coordination or approval to create or modify. Plugins consist of user-defined, sharable XML resource descriptions which can be installed into individual user accounts; the user interface uses information contained in these descriptions – such as link schemas – to send appropriate data such as textual object references to specified external resources, which can then request full-text data, metadata, and other relevant content via API and perform task-specific processing on the requested data. Any user can create a new plugin, share it with others, and take responsibility for future updates to their plugin code, without requiring central approval or coordination.

This technical framework enables a distributed web-based development model in which external projects can be loosely integrated with the digital library and its user interface, from an end user perspective being well integrated with the library, while from a technical standpoint being developed and maintained entirely independently. Currently available applications using this approach include simple plugins for basic functionality such as full-text export, the "Text Tools" plugin for textual analysis, and the "MARKUS" named entity markup interface for historical Chinese texts developed by Brent Ho and Hilde De Weerdt, as well as a large number of external online dictionaries. The "Text Tools" plugin provides a range of common text processing services

and visualization methods, such as n-gram statistics, similarity comparisons of textual materials based on n-gram shingling, and regular expression search and replace, along with network graph, word cloud, and chart visualizations; “MARKUS” uses external databases of Chinese named entities together with a custom interface to mark-up texts for further analysis. Because of the standardization of format imposed by the API layer, such plugins have access not only to structured metadata about texts and editions, but also to structural information about the text itself, such as data on divisions of texts into individual chapters and paragraphs. For example, in the case of the “Text Tools” plugin this information can be used by the user to aggregate regular expression results and perform similarity comparisons by text, by chapter or by paragraph, in the latter two cases also making possible visualization of results using the integrated network graphing tool. As these tasks are facilitated by API, tools such as these can be developed and maintained without requiring knowledge of or access to the digital library’s code base or internal data structures; from an end user perspective, these plugins do not require technical knowledge to use, and can be accessed as direct extensions to the primary user interface. This distributed model of development has the potential to greatly expand the range of available features and use cases of this and other digital libraries, by providing a practical separation of concerns of data and metadata creation and curation on the one hand, and text mining, markup, visualization, and other tasks on the other, while simultaneously allowing this technical division to remain largely transparent to a user of these separately maintained and developed tools and platforms.

Surviving the storm: The Revival of Edo Cultural Traditions through Local and Cultural Networking

Kumiko McDowell, Kevin McDowell (University of Oregon Knight Library)

In the late 19th century the practice of pasting nosatsu, wood block votive slips, on shrines and temples, became popular among commoners in Edo. A number of nosatsu-kai, that is groups for nosastu pasting and exchanging emerged. Each member commissioned a team of artisans including a designer, a carver and a printer to create original multi-colored nosatsu slips for exchange gatherings and competed to see who could produce the highest quality of nosatsu.

Nosatsu-kai and nosatsu practice, which have traces of the Edo culture from the previous era largely declined in the early Meiji period mainly because of changes caused by the Meiji Restoration, such as Westernization and modernization throughout Japanese politics and society, negation of Edo culture and tradition, and haibutsu-kishaku that is a movement to abolish Buddhism and Buddhist temples.

However, in the mid Meiji era, nosatsu-kai revived and gradually expanded. They regularly hosted nosatsu exchange meetings as well as joint meetings. As a historical and social background of the nosatsu revival and development in the Meiji era, the following aspects that brought people re-evaluate and appreciate nostalgic conventional culture from the previous period should be considered: a tendency in the middle of the Meiji period, at least, partly to reject Western culture and look to traditional forms of Japanese art and entertainment. The 300 year anniversary of Tokugawa Ieyasu's founding of Edo was celebrated in 1889, showing that some segments of Japanese society were harkening back to Edo period culture. In addition, due to the Sino-Japanese War (1894), nationalistic fervor swept through the nation.

With these background phenomena, more importantly, local and social network enhanced nosatsu-kai revival. In the Edo period commoners were assigned to live in certain areas depending on their social class. For example, Nihonbashi was a merchant town and Kanda and Fukagawa were artisans' neighborhoods. They established local communities within their areas. In addition, social networks developed through the cultural activities, including poetry groups such as haikai and kyoka groups. These cultural groups expanded from the upper class to the middle class with the economic prosperity in the Edo period that allowed commoners to have access to arts or poetry. After the Meiji restoration, some of these traditional networks survived and more developed since they were from the conventional classes in Japanese society. Among these network groups, I suggest that the most influential ones for the nosatsu-kai revival were these three groups: Kanda and Nihonbashi local areas, Shukokai/antique collectors, and Choyukai (Tatoo lovers group) through examining the members of revived nosatsu-kai and designs often seen in nosatsu from the era.

The Kanda and Nihonbashi neighborhoods, areas located in the center of Tokyo, were largely populated by merchants and artisans during the Meiji and Taisho eras. The Kanda/Nihonbashi local area network includes two groups: Parishioners of Kanda Myojin and workers at fruit and vegetable markets in Kanda. Kanda Myojin is the local tutelary shrine for the areas. Devoted parishioners organized two groups called Miyakagi-ko 宮鍵講 and Ofusegi-ko御防講 to serve and protect Kanda Myojin. There used to be huge fruit and vegetable markets in Kanda since the Edo period. Merchants who did their business in the market usually both owned a shop in the area also resided there. Therefore, there was a strong sense of community in the Kanda market area.

Members of the Shukokai (集古会)/antique collector group were interested in collecting old things including toys, books and prints from the previous period. The Shukokai was established in 1896, and the group later divided into sub-groups depending on their types of collections and held regular meetings to show their collections and exchange ideas. The members came from

various social classes including noblemen, intellectuals, politicians, and commoners.

The Choyukai is a group devoted to tattooing and was originally established in Kanda by Iseman and other people who had tattoos designed by Horiuno (彫 宇 之).The members joined shrine festivals including the Kanda Myojin festival and the Fukagawa Hachiman festival, to carry portable shrines.

I selected major members of nosatsu-kai (1902-1920) including Iseman (いせ万) , Takahashito (高橋藤), Konsan (紺三), Chikabo (千加坊), Setcho (櫛朝), Izuaka (いづ赤), Shimayone (しま米), Kiwashichi (キ 0 七), Senrei (扇令), Tataume (田々梅), Daigin (大銀), and Futami (二見) to show their networking group and relationship and found that, Iseman was at the center of all three networks among these nosatsu enthusiasts. He was an owner of a vegetable and fruit wholesale shop in Kanda, commissioned large numbers of gorgeous exchange nosatsu cards throughout his life, hosted many sessions of nosatsu-kai meetings. He was an antique collector and was one of the founders of the Choyukai. He was the key person for the nosatsu-kai network and his networking played an important role for the revival of nosatsu pasting and exchanging activities in the Meiji and Taisho periods.

I'm thinking of employing digital tools to show the networks of nosatsukai-members and relationship between them by area and each group. Palladio is the best candidate for now. This project is the first one to analyze and visually describe the nosatsu-kai members and their cultural and regional relationships in late Meiji and Taisho in terms of leaders. I am also thinking of applying a digital tool for related projects. First, I will include more nosatsu-kai members from the period to examine the possibility of the existence of other networks/relationships. Next, I will study the influence of Dr. Frederick Starr, who was an American anthropologist and devoted nosatsu enthusiast, on the development of nosatsu-kai in Taisho era in terms of his networking in Japan. Moreover, I am thinking about comparing cultural networks in the Edo and Meiji era. I believe that mapping software can be a great tool to demonstrate how cultural networks consist and develop.

Towards a Conceptual Framework for Superworks

Senan Kiryakos, Shigeo Sugimoto (University of Tsukuba)
Jacob Jett, J. Stephen Downie, Yi-Yun Cheng (University of Illinois at Urbana-Champaign)
Jin Ha Lee (University of Washington)

Introduction & Context

While the investigation of traditional cultural products like text remains the staple of humanities research, there is a growing group of humanists (both digital and traditional) who are focusing on more recent, popular culture (pop-culture) artifacts and works (e.g., Kashtan, 2011; Helms, 2015; Losh, 2015; Whitson & Salter, 2015). These pop-culture artifacts are undeniably an important part of our cultural heritage; recent statistics show that the comics market in the U.S. alone has reached 1.03 billion dollars (Reid, 2016) and consumer spending in video game industry in the U.S. is estimated at 23.5 billion dollars (ESA, 2016). Pop-culture artifacts are an integral part of today's society and their study by scholars is informative with regards to determining what kinds of things society values and how artists, authors, and other cultural content producers interact with and remark upon those values.

Many of these works break the boundaries of traditional humanities research through being produced using multiple media formats. Born-digital entities like video games are an example multimedia that digital humanists studying pop-culture (Rockwell & Mactavish, 2004) engage with. Serial narratives spanning multiple formats like the *.hack// series* (Jett et al., 2017), and entities that spawn multiple, parallel series, such as *Gundam*, *Ghost in the Shell*, and the *Marvel Multiverse*, form more complex examples.

Complex Work-like Entities ("Superworks")

The role of adequately modeling entities in information retrieval (IR) systems that digital humanists use is well recorded (McCarty, 2004) and efforts like Walsh's Comic Book Markup Language (2012) provide a fine-grained, TEI-like approach. But many IR systems are based on descriptive metadata rather than richly annotated data. This poster seeks to examine the boundaries of this more complex entity via the notion of "superwork" in the pop-culture space, the justification for which has been noted by others previously (McDonough et al., 2010), as these entities regularly span numerous series and mediums. We provide an example of *Gundam*, hoping that this approach can help provide an opportunity for humanities scholars to reflect their views and opinions in the design of a conceptual model for this more complex entity.

Gundam as a Superwork Exemplar

Looking at the pop-culture phenomenon *Gundam* in aggregate, we see a conceptual entity spanning multiple mediums, such as manga, animated films and TV series, games, etc. This umbrella entity called *Gundam* goes beyond the aspects of the FRBR Work concept that deal with formats. There is a lack of understanding about what information aggregates made up of related resources are describing, as there is no entity to represent this level in isolation from format-specific information. For instance, consider a selected portion of the *[Original] Gundam Series* in Figure 1 (below).

For serial works in general, sub-works that form parts are related by being parts of wholes, which are themselves parts in bigger wholes (IFLA 1998, p 29). What is unclear is how the alternate formats (i.e., alternate editions) of the core work relate to the overall series. In their paper describing the relationships that exist among video games, Lee et al. (2014) develop the "isRemakeOf" relationship (p 6), which articulates exactly how the novel series and anime film series in Figure 1 relate to the anime series. However, it is still unclear how the "remakes" are related to the sequel and prequel manga.

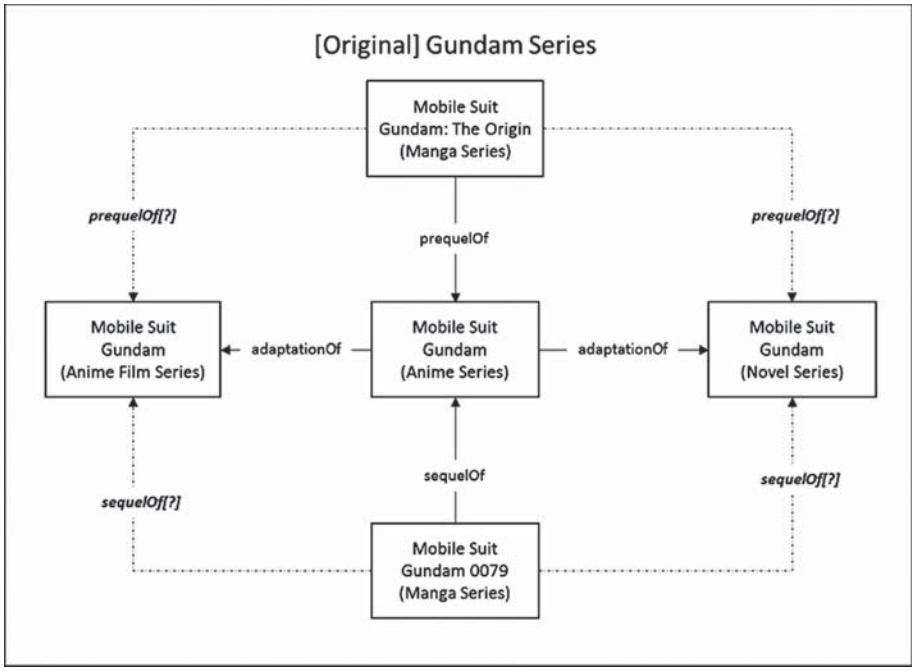


Figure 1: Work Model of portion of the *[Original] Gundam Series*

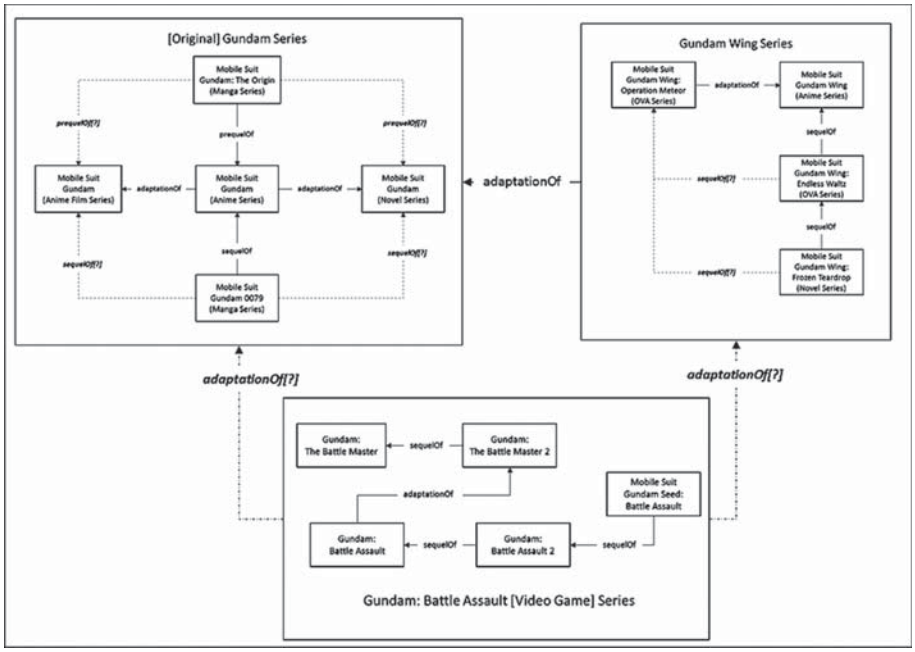


Figure 2: Uncertain Relationship Model Between [Partial] Serial Works

Figure 2 showcases how this problem expands when an entire ecosystem of linked works is brought together. From a human perspective, the video games in the *Gundam: Battle Assault [Video Game] Series* are related to the *[Original] Gundam Series* and the *Gundam Wing Series* but the nature of the relationship is not easily articulated by a conceptual model and thereby the series are not easily retrieved together in an IR system setting. Together these three series entities form a large information aggregate and, the creation of a "superwork" entity that represents information aggregates like the one in Figure 2 can help solve this issue. Our future efforts will include closely examining the "superwork" entity concept to better understand properties that are unique to it.

References

[1] ESA (Entertainment Software Association). (2016). Essential Facts about the Computer and Video Game Industry. Retrieved from: http://essentialfacts.theesa.com/Essential-Facts-2016.pdf

[2] IFLA Study Group on FRBR (IFLA). (1998). *Functional requirements for bibliographic records: Final report [revised 2009]*. München: K.G. Saur Verlag.

[3] Jett, J., Humpal, N., Valentine, C., & Lee, J. H. (2017). What is a series, really? *Knowledge Organization 44*(1), pp 24-36.

[4] Kashton, A. (2011). Forward to the past: Nostalgia for handwriting in *Scribblenauts* and *The World Ends with You*. *Digital Humanities Quarterly 5*(3). Retrieved from: http://digitalhumanities.org:8081/dhq/vol/5/3/000098/000098.html

[5] Losh, E. (2015). What can the digital humanities learn from feminist game studies? *Digital Humanities Quarterly 9*(2). Retrieved from: http://digitalhumanities.org:8081/dhq/vol/9/2/000200/000200.html

[6] Lee, J. H., Clark, R. I., Sacchi, S., & Jett, J. (2014). Relationships among video games: Existing standards and new definitions. Paper presented at the *77th ASIS&T Annual Meeting*, 31 Oct.-5 Nov. 2014, Seattle, WA, USA.

[7] McCarty, W. (2004). Modeling: A study in words and meanings. In S. Schreibman, R. Siemens, & J. Unsworth (eds.) *A Companion to Digital Humanities*. Retrieved from: http://digitalhumanities.org:3030/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-3-7&toc.depth=1&toc.id=ss1-3-7&brand=9781405103213_brand

[8] McDonough, J., Kirschenbaum, M., Reside, D., Fraistat, N., & Jerz, D. (2010). Twisty little passages almost all alike: Applying the FRBR model to a classic computer game. *Digital Humanities Quarterly 4*(2). Retrieved from: http://digitalhumanities.org:8081/dhq/vol/4/2/000089/000089.html

[9] Reid, C. (2016).North American Comics Market Reaches $1.03 Billion. Publishers Weekly. Retrieved from: http://www.publishersweekly.com/pw/by-topic/industry-news/comics/article/70897-north-american-comics-graphic-novel-market-reaches-1-03-billion.html

[10] Rockwell, G. & Mactavish, A. (2004). Multimedia. In S. Schreibman, R. Siemens, & J. Unsworth (eds.) *A Companion to Digital Humanities*. Retrieved from: http://digitalhumanities.org:3030/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-2-10&toc.depth=1&toc.id=ss1-2-10&brand=9781405103213_brand

[11] Walsh, J. A. (2012). Comic book markup language: An introduction and rationale. *Digital Humanities Quarterly 6*(1). Retrieved from: http://digitalhumanities.org:8081/dhq/vol/6/1/000117/000117.html

[12] Whitson, R. T. & Salter, A. (2015). Introduction: Comics and the digital humanities. *Digital Humanities Quarterly 9*(4). Retrieved from: http://digitalhumanities.org:8081/dhq/vol/9/4/000210/000210.html

Prototype of Linked Open Data Model for Tang Poems

Yan Cong, Masao Takaku (University of Tsukuba)

In recent years, linked open data (LOD) technology has emerged, and many cultural institutions use it for making their information freely available to the public. This paper focuses on information of Tang poem cultural resources, which were written during the Tang Dynasty of China from the 7th to the 10th century and proposes a data model along with its prototype for Tang poems as LOD. Tang poetry is one of the indispensable information resources in classical literature. It has been translated in several languages to provide people with the opportunity to study and use it. When students are being taught Tang poetry in Japanese, it is necessary to deal with a characteristic style of expressions in Japanese, which is quite different from other languages. Our final goal is to provide an application that consists of an LOD dataset with rich semantics, the fulltext data with various expressions, and a cross-lingual information retrieval system in Chinese, Japanese and English. Not only can this help people learning Tang poem, but it can also assist Tang poem teaching and teacher education in different languages.

In this paper, we especially aim to collect and organize information about Tang poems, which are used in textbooks. Moreover, we construct a structured data model which covers Tang poem works and instances, different content expressions of Tang poems and the relationships with their textbooks. The metadata of Tang poems includes author name, title, and the four types of styles which exposit the Tang poem styles in two features. One feature is the length of line with 5 and 7 syllables, and the other one consists of jueju (絶句) and lüshi (律詩).

For the data model, we employ the BIBFRAME data model, a standard bibliographic framework developed by the Library of Congress. The basic structure of BIBFRAME consists of work and instance. Tang poems are modeled using two separate models: an original and unique poem is regarded as a work entity, and a publication of a poem (work) is regarded as an instance. For the properties, we utilize several standard vocabularies such as BIBFRAME, Schema.org, and Dublin Core, as well as our own vocabulary for Tang poems.

As a dataset, we used Tang poems included in Japanese textbooks. We collected 374 Tang poems as instance entities. These instances consisted of 59 unique poems, which were modeled as work entities written by 25 authors, and all poems were retrieved from 53 textbooks used in Japanese junior high schools and high schools, as of 2016. We named four types of content expressions as unpunctuated text (hakubun; 白 文), punctuated text (kundokubun; 訓 読 文), reading text (kakikudashibun; 書き下し文), and translation text (honyakubun; 翻訳文). Punctuated text means text which transcribes Chinese classic text with markup annotation. It is a method of annotating classical Chinese so that the text can be easily read in Japanese. Not only the Tang poems, but also most of the Chinese classic texts were transcribed in these styles. Figure 1 shows examples of punctuated and reading expressions of the same Tang poem “Thoughts in a tranquil night” by Li Bai.

In order to design the LOD dataset, we defined several kinds of URIs for Tang poems as follows: We identified 59 unique poems as works, and defined a resource URI as https://w3id. org/ tangpoem/*n*, where *n*, *n* = 1, 2, ..., 59, corresponds to a work identifier. We identified 374 tang poems as instances, and defined a URI as https://w3id.org/tangpoem/instance/*m*, where *m*, *m* = 1, 2, ..., 374, symbolizes an instance identifier. The four types of Tang poem styles have been defined as https://w3id.org/tangpoem/style/*style_name*, where *style_name* corresponds to the name of styles 5, 7, 絶句 and 律詩 . Furthermore, a Tang poem author is defined as https://w3id. org/tangpoem/author/*name*, where *name* is the name of the author. Full text content of a Tang poem is defined as https://w3id.org/tangpoem/instance/*m*/ content, where *m* means the identifier for the instance.

Using the LOD vocabulary and the collected datasets, we built a Tang poem data model. In this structured model, a Tang poem work is linked to all its instances, and then the instances are

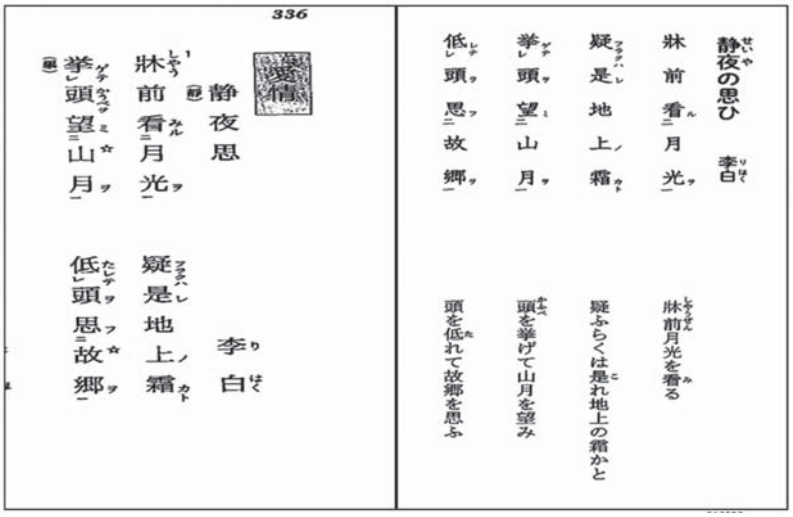


Figure 1: Comparison between different expressions of a Tang poem[1][2]

linked to, as much as possible, related information, including the textbooks or other data. Figure 2 illustrates an excerpt of the structural model for Tang poem 1. In the model, there are several kinds of relationships in respect to Tang poem 1. Tang poem work 1 is realized as instance 8, which is contained in a junior high school textbook [2]. Instance 8 has a property bf:instanceTitle, which is the name of the title used in the textbook, and a property bf:partOf, which is the relationship between instance 8 and the textbook. A full text content of Tang poem instance 8 is modeled as the type of reading text, and instance 8 has a property tangpoem:content, which refers to full text content for the instance, and a property tangpoem:fulltextType, which means that the full text content's type is that of reading text.

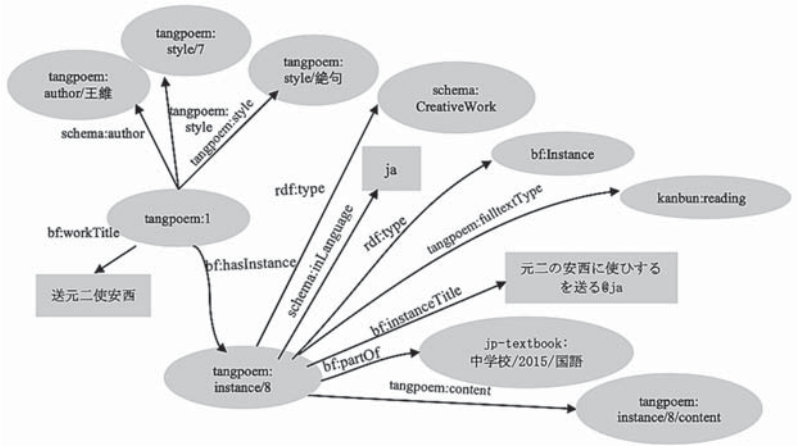


Figure 2: The structured model for tangpoem:1

This paper described an approach to model Tang poems in the textbooks of 2016 using LOD, and discussed the issue of describing a context of Tang poem special expressions in Japanese. In future research, we will attempt to 1) find an appropriate way to identify and name Tang poem works with interoperable identifier schema, 2) encode expressions of full text content with Text Encoding Initiative, and 3) build an application of Tang poems with LOD.

Acknowledgment

This work was partially supported by JSPS KAKENHI Grant Number JP16H02913.

References

[1] Katsumi Togo, et al. 2013. “Koutou Gakkou Kokugo Sougou (高等学校国語総合)”, Daiichi Gakushu Inc. https://w3id.org/jp-textbook/ 高等学校 /2012/ 国総 /326
[2] Jun’ya Noji, et al. 2015. “Chuugakou Kokugo 3 (中学校国語 3)”, Gakko Tosho Inc. https://w3id.org/jp-textbook/ 中学校 /2015/ 国語 /928

Extracting the Patterns of Harmonic Features within Mozart's Symphonies and String Quartets Compositions based on the Notion of Pitch-Class Set

Michiru Hirano, Hilofumi Yamamoto (Tokyo Institute of Technology)

Introduction

The study seeks to demonstrate how the notion of pitch-class set (Forte 1973) is useful for describing the harmonic features within Mozart's symphonies and string quartet compositions. It has been proposed that there are differences between the scores for orchestral and chamber compositions, that reflect divergences in terms of the numbers of string players (Hickman 1981). This is- sue has previously addressed by focusing on Mozart's symphonies and string quartet compositions, particularly in terms of the relationships between parts (Hirano and Yamamoto 2016a,b).

In this study, we focus on the set of pitches that appear across a complete part and compare the two genres in terms of their harmonic features. Although differences in the harmonic features of orchestral and chamber compositions have been proposed, such that harmony is generally simple with occasional dissonances for orchestral pieces while there is greater use of dissonance and more daring harmonic shifts for chamber compositions (Hickman 1981: Table 1), those differences have yet to be demonstrated objectively.

Pitch-class set (pc set) refers to the set of distinct integers that represent the pitch classes (Forte 1973: p.3). We utilize pc sets to summarize the various arrangements of notes on a score and to represent them as minor patterns. Although pc set is unsuitable for describing the functions of chords, unlike classical harmonic theory, it can rep- resent harmonic features to some extent, but, most of all, its major advantages lie in avoiding researcher subjectivity and in being able to automatically process large datasets.

Method

Materials

We target Mozart's 59 initial movements, of which 39 are symphonies and 20 are string quartet compositions. The scores were con- verted into MusicXML, which is a machine readable format.

Data Structure

The pc set for every measure within the materials was determined as follows (Forte 1973).

1. Each pitch that corresponds to one of 12 distinct pitch classes is replaced by the integers 0 to 11.
2. All the pitches that appear within a measure together form the pc set, with repetitions eliminated.
3. The pc sets are sorted according to the normal order, based on the least difference determined by subtracting the first integer from the last.
4. The pc sets are transposed to the prime form, which is the form where the first integer is 0.

Through this procedure, a measure containing a major triad (such as a chord constructed from C, E and G) would, for in- stance, be represented with the notation of [0, 4, 7], regardless of the pitch of its root.

Results

Theoretically, there are 351 patterns of pc sets. Of those, Mozart used 270 patterns (77%) within the initial movements of the analyzed symphonies and string quartet compositions.

We examined the relationship between the length of a work and the number of pc sets used within the work (Fig. 1). Given that the overall correlation coefficient is 0.84, a strong correlation would appear to exist between length and number of pc sets.

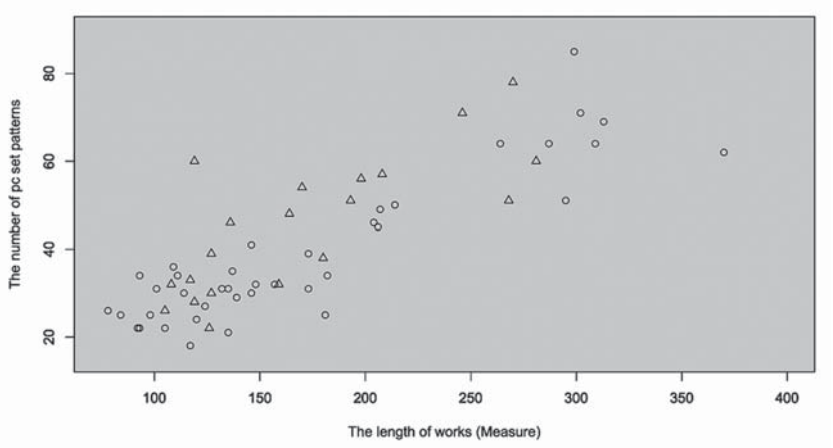


Figure 1: The relationship between the length of a work and the number of pc sets used within the work. The horizontal axis represents the total number of measures. Circles and triangles represent symphonies and string quartet compositions, respectively.

However, the cluster of works to the right side of the scatter plot are more randomly disperse than the larger cluster to the left, and, indeed, the correlation coefficient for those 12 works is -0.03.

Figure 2 shows the increases in the number of pc sets along a time axis for the K.551 symphony K.551 and the K.590 string quartet composition, which are the last examples of these respective genres. Both symphonies and string quartet compositions are constructed from a sonata form, which consists of three sections: (1) exposition, (2) development and (3) recapitulation. Table 1 presents the numbers of new pc sets added across the sonata sections of the two example compositions.

Table 1: The numbers of new pc sets added across the sonata sections of the two example compositions.

Composition	(1) Exposition	(2) Development	(3) Recapitulation
Symphony K.551	43	14	12
String quartet K.590	36	13	7

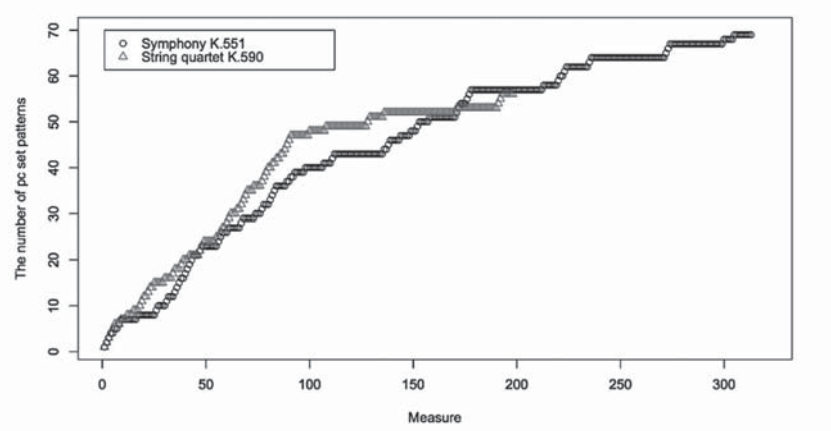


Figure 2: The increases in the number of pc sets along the time axis. The horizontal axis represents the time series based on the measures.

Discussion

The present results indicate that Mozart did not utilize all the theoretical patterns of pc sets throughout his symphony and string quartet compositions. Although the number of patterns employed within a work is highly depend on the length of the work, there is no correlation between length and number of sets in the case of the 12 works that have more than about 250 measures.

Table 1 indicates that there is a correspondence between the introduction of new patterns and sonata section. It is natural for new patterns to be introduced during the earlier exposition and development sections. More interestingly, new patterns are also introduced during the final recapitulation section, which suggests that the recapitulation section is not merely a copy of the exposition section.

The findings from the present study suggest that pc sets could be useful for future studies that seek to capture the differences between symphonies and string quartet compositions.

Conclusion

The present study examined the variety of pc sets used by Mozart within his symphonies and string quartet compositions, and demonstrated that the method is potentially useful for summarizing musical scores.

References

[1] Forte, Allen (1973) The Structure of Atonal Music: Yale University Press.
[2] Hickman, Roger (1981) “The Nascent Viennese String Quartet,” The Musical Quarterly, Vol. 67, No. 2, pp. 193-212, Apr.
[3] Hirano, Michiru and Hilofumi Yamamoto (2016a) “Instrumentation of Mozart’s Symphonies: Study for Width and Thick- ness of Musical Parts,” in Proceedings of Jinmoncom 2016, pp. 63–68.
[4] Hirano, Michiru and Hilofumi Yamamoto (2016b) “Quantitative Analysis for Mozart’s Musical Style: The Performance Form of Eine kleine Nachtmusik.,” in Proceedings of AEARU Young Researchers International Conference, pp. 47–50.

Creating A Work Entity Dataset of Console Games Using Wikipedia

Tetsuya Mihara, Mistuharu Nagamori,
Shigeo Sugimoto (University of Tsukuba)

Background

Manga, animation, and video games are known as representative new media of Japanese pop culture. Creations in those different media types have close relationships with each other, e.g. expressing the same story, or sharing the same universe or characters. These relationships are useful for users to search them. The authorized dataset of these media and their relationships is required by not only users but also researcher including those who are interested in digital humanities for new media. It helps to understand the dynamics of new media creations.

However, reliable resources about relationships of creations are insufficient, as archiving these media types is a relatively new activity. It is difficult to create the authorized dataset from scratch since thousands of creations had been already produced and new creations are producing every day. Media Art Database (Development Version) (MADB) [1] is a database of manga, animation, and video game published in Japan, produced by Agency for Cultural Affairs in Japan as the national authority of these creation. However, MADB is created from records of packages of each medium, and it lacks the information about relations between each medium. On the other hand, the information of these relations is available from some resources created by users, e.g. Wikipedia. Unfortunately, these resources are not suitable to search and reuse.

In this research, we try to create data of the relationships between package records of manga, animation, and video games using resources from the Web. This paper presents the method and an experiment creating Work entities of console games from Wikipedia.

Research Objective

Users of console games need conceptual entities to describe a whole creation or groups of creations, similar to other media types such as books. Therefore, this research defines two entity types for the dataset of console games, "package entity" and "work entity". Package entity represents a publication recognized by titles and specific game console. Work represents a concept to embody package entities. This idea is based on FRBR [2], which is a conceptual model of the bibliographic records to describe entities, relationships, and attributes in online library catalogues and bibliographic databases from a user's perspective.

We focus on describing three characteristics of console game publication by the relationships between the Work entity and Package entity.

Characteristic 1: Porting

Some video games for a certain game console are reproduced for different game consoles. This is called "Porting". Users need the information about porting of the video game when they want to know whether they can get the edition for the game console they have.

Characteristic 2: Series

Some game titles compose a series of the game or belong the same game franchise. These groups are used frequently to find games.

Characteristic 3: Transmediation

Some games also have close relationships with other creative works of deferent medium, e.g.

animation and manga. Users need not only records of game titles but also these related works because they are helpful to understand the whole stories and universes of the game.

How to Extract Work Entities of Console Games in Japan from Wikipedia

The information about the three characteristics shown above is contained in articles on Wikipedia about video games. We try to extract work entities of console games in Japan from articles in Japanese Wikipedia [3]. The extracting method consists of three steps,

Step 1. Getting the package resources from the list

Lists of game package are available articles belonging the category "Video game lists by platform" (「ゲーム機別ゲームタイトル一覧」in Japanese). The lists are ordered by platforms and published year. A record in the list contains texts of title and publish date of each video game packages.

Step 2. Finding work-instance relationships from the links to article

Some records contain a link to the article related to the video games. What these linked articles are written about is various. Sometime it is about gameplay, series, or group of creations by user's needs and length of the article. Some links from different records to a same article. These links represent that these records are belonging same group of a creation . In other words, These links represent work-package relationships about the three characteristics.

Step 3. Identifying media type and series from the DBpedia

DBpedia is Linked Open Data dataset created from Wikipedia. In DBpedia, the article of Wikipedia is defined as a RDF resource, and information of the article (links, categories, templates) is provided as properties of the RDF resource. The RDF resource of DBpedia often has some properties to describe its media type and group of series. We considered these RDF resources as the one of work entity described by the article and recognized information about transmediation (especially manga and animation) and series from the properties.

Experimental Results of the Extraction

We experimentally extracted 26,003 package resources and 17,966 links to the related article from "Video game lists by platform" in Wikipedia, and 8,199 work entities is created from 17,966 articles except duplication of links. 3,770 works aggregate several package entities, 2,414 works are grouped by series, and 1,736 works have relationships with manga and animation creation.

References

[1] Media Art Database (Development Version), https://mediaarts-db.bunka.go.jp/
[2] FRBR, https://www.ifla.org/publications/functional-requirements-for-bibliographic-records
[3] Japanese Wikipedia, https://ja.wikipedia.org/wiki/

Visualizing Narratives of the Atomic Bombings of Hiroshima and Nagasaki

Steven Braun (Northeastern University Libraries), Kelsey Menninga

In the book *Narrative*, Paul Cobley begins by emphasizing the primitive role of storytelling in human history, stating "...humans have a compulsion to narrate. Especially after the development of the verbal faculty, human beings have constantly told stories, presented events and squeezed aspects of the world into narrative form. Wherever there are humans there appear to be narratives." Indeed, narratives operate at many different levels of our daily lives and take on many forms. There is written narrative, which is embodied in the many forms of literature and the novel; oral narrative, represented by the sharing of stories passed down through cultures and generations; and historical narrative, the retelling of historical events through stories spatio-temporally situated in people and places. And then there is a new kind of narrative that has emerged with particular significance in recent decades: visual narrative, or the retelling of events and stories through images, including cinema and photography. In this sense, data visualization – or information design more generally – is another kind of visual narrative, one in which the act of retelling is founded upon data and the grammar of that retelling is derived from points, lines, and planes.

Data visualization is thus a powerful medium for representing the manifold nature of narrative expressed in other forms. By acknowledging the constructedness of knowledge and its representation, data visualization invites opportunities to recognize the ways in which the experiences and identities of both user and designer alike are fundamentally implicated in the creative and interpretative acts. For this very reason, visualization can thus also operate as a medium for reconciliation, honoring the complexities of narrative when understood as an intersubjective act between designer and user.

In this project, we seek to explore this role of visualization as reconciliatory medium through the lens of two specific events in history: the dropping of the atomic bombs on Hiroshima and Nagasaki, Japan. Focusing on narratives of these events that are presented in history textbooks in the United States and Japan, we wish to explore ways in which we may use visualization to represent and communicate the manifold retellings of these events, their parallels, and their differences. While similar work has been done previously through the *Divided Memories and Reconciliation* project at Stanford University (http://aparc.fsi.stanford.edu/research/divided_memories_and_reconciliation), this research seeks to extend that prior work in two primary ways. First, we endeavor to collect a larger sample of narratives from textbooks in both countries, providing a database of information that scholars may use for further research. Second, we intend to present those data in a visual form, thereby increasing their accessibility to wider audiences. The hope is that by providing greater access to these narratives, new paths toward reconciliation between the United States — "thick" reconciliation, driven by the people — can be built. In depicting many different narratives from United States and Japanese textbook portrayals of these events, we hope to demonstrate that there are many different ways of observing the same events in history, each of which may have its own strengths and faults.

In the process, we hope to explore the following questions:

- What does it mean to think of data visualization as narrative, and how does that inform the ways in which we engage with visualization as a medium for communication?
- How can visualization be used to guide critical inquiry around events in history that may be highly emotionally and politically charged in their reception?
- What can visualization reveal about how these events in history have been differentially depicted, and how do those differences affect how we engage with those events on practical, societal, and cultural levels?

To address these questions, we will carry out this research in several stages. In the first stage, we will collect as many different narratives of the dropping of the atomic bombs on Hiroshima and Nagasaki as possible from different textbooks. In this exploratory phase, we are limiting our analysis to textbooks specifically due to the known biases and controversies associated with them in both America and Asia, making them particularly relevant for historiographic study. For our purposes, "textbook" refers to any primary source used in history courses, and the temporal scoping of "dropping of the atomic bombs" will be determined by the contextual events immediately preceding and following the bombings that are depicted in the texts. Importantly, to enhance feasibility, data collection will be limited to sources from the United States and Japan only, and data may be collected through crowdsourcing efforts that engage the participation of the general public. In the second stage, Japanese texts will be translated into English, and both translated and English-origin texts will be encoded for visualization. In the final stage, those texts and data will be visualized and shared via an interactive exhibit over the internet.

As this project is in its early stages of development, this poster will describe a prototypical approach to these analyses and possible platforms for data collection, including platforms that engage the general public, as well as important considerations for further research such as international copyright law. It will also present prototypes for potential visualizations, *e.g.*, interactive and static alternative representations of linear and circular timelines that enable cross-comparison of multiple textbook narratives. As this research involves the collection and study of highly unstructured textual data (*i.e.,* textbook content), this poster will also discuss pitfalls and considerations in working with such data for the purposes of visualization.

Deep Features for Image Classification and Image Similarity Perception

Zhenao Wei, Lilang Xiong, Kazuki Mori, Tung Duc Nguyen, Tomohiro Harada, Ruck Thawonmas, Keiko Suzuki, Masaaki Kidachi (Ritsumeikan University)

***Abstracts* - This paper describes a study that examines deep features in image classification and image similarity perception. This work is inspired by previous findings in image style classification that correlation between a deep-learning network's feature maps can effectively describe image features for classification. Both objective and subjective experiments are conducted to evaluate existing methods that derive a feature vector for a given image from feature maps in a deep learning network.**

Keywords - deep features, ukiyo-e, Gram matrix, feature vector, cosine similarity, similarity perception

Introduction

In this paper, we introduce our research achievements in extracting deep features of images by using a convolutional neural network called VGG-19 [1]. The effectiveness of extracted deep features are evaluated on two tasks, image classification and similarity perception, which underlie our final goal of building a recommendation system for *ukiyo-e* images in a database[1] of the Art Research Center (ARC), Ritsumeikan University. In particular, we focus on the most popular 5 genres: *yakusha-e, bijin-ga, doban-e, meisho-e*, and *monogatari-e* (prints of actors, beauties, landscapes, and stories, and copperplate prints).

Previous Work

Chu and Wu [2] and Matsuo and Yanai [3] independently reported that VGG-19's conv5_1 layer, having 512 feature maps, is the most effective layer in image style classification. In Chu and Wu, many types of feature vectors were investigated, but, due to limited space, three of them are described as follows:

(1) Gram vector: This vector is constructed by the diagonal elements and their lower-part elements in the Gram matrix of conv5_1. Each element, out of 131328 (512 × (512 + 1)/2) elements, is the dot product of the corresponding pair of feature maps.
(2) Cosine-sim vector: Rather than the dot product, cosine similarity is computed.
(3) Gram-dot-cosine vector: This vector is the result of element-wise multiplying the two vectors above.

According to Chu and Wu, the Gram-dot-cosine vector is the most effective among all feature vectors that they examined.

Matsuo and Yanai reported in their work that the following feature vector from conv5_1 was the most effective.

(4) Gram-sgnsqrt vector: This vector is the same as the Gram vector, but with each element being L2 normalized with signed square root.

[1] http://www.dh-jac.net/db/nishikie-e/search.php?enter=default

The number of dimensions of each of the four feature vectors above was then reduced, e.g., to 4096, using principal components analysis (PCA) in their experiments.

Table 1: Classification results of ukiyo-e classification by SVM

Feature Vector	Gram	Cosine-sim.	Gram-dot-cosine	Gram-sgnsqrt	Content
Classification Rate	86.65 %	**88.56 %**	86.11 %	80.76 %	84.46 %

Figure 1: An example set in the user study

Experiments

We conducted two experiments using 10,000 *ukiyo-e* images randomly sampled from the ARC database such that each of the five genres equally has 2000 images. The first experiment is image classification where feature vectors of 4096 dimensions derived first by each of the four methods and then reduced by PCA are used as input to a support vector machine (SVM). For reference, we also experimented with another type of feature vector (called Content vector) resulting from directly applying PCA to a feature vector whose elements are the outputs of all feature maps in conv5_1. Five-fold cross-validation was employed in this experiment.

Table 1 shows the classification results of the SVM with each of the five feature vectors as input. It is interesting to see that, in contrary to previous work, for Ukiyo-e images, the cosine-sim vector is the most effective. Since the Gram-sgnsqrt vector and Content vector are the least effective for Ukiyo-e, we removed them from our user study described below.

In the second experiment, 15 images (three per genre) were randomly selected as queries. Twenty college students participated in this user study. Two measures were used to find the most similar image to a given query by each of the three remaining features. One is the Euclidean distance (finding the image with the lowest distance to the query) and the other is the cosine similarity (finding the image with the highest similarity to the query). To test the participants' reliability, two test sets were added whose query image was also used as a candidate image.

Each participant was shown, in a random order, 15 sets of a query image and its similar images according to each of the three features, and was then asked to vote the image that she or he thought most similar to the query image (e.g., Fig. 1). In the experiment, resulting images that are identical to a query image but with different file names were removed and replaced by the next most similar candidate images. However, we regarded images having the same content but with different colors as different. Please also note that if resulting similar images are the same by different features, only one image will be shown; and if it is voted, this vote will be counted for all corresponding features.

All participants correctly selected the image identical to the query image in the aforementioned two test sets, and Table 2 shows the results of this user study where G, C, and GdC denote the Gram, Cosine-sim, and Gram-dot-cosine, respectively, with “e” and “c” the Euclid distance and the cosine similarity. These results also show that the cosine-sim feature outperforms the other features, in particular, when the cosine similarity is used as the measure.

Table 2: Results of the user study on ukiyo-e similarity perception

Feature-Measure	G-c	C-c	GdC-c	G-e	C-e	GdC-e
# of Votes	78	**121**	65	54	108	71

Conclusions and Future Work

Our main finding is that for *ukiyo-e* images, the best feature extraction method, in terms of both image classification and image similarity perception, is the one that generates a feature vector whose element is cosine similarity between the corresponding pair of feature maps. As our future work, we will develop a content-based recommender system for *ukiyo-e* images based on features extracted by this method.

References

[1] L. A. Gatys, A. S. Ecker, and M. Bethge. Image Style Transfer using Convolutional Neural Networks. CVPR 2016, pp. 2414-2423.
[2] W. T. Chu and Y. L. Wu. Deep Correlation Features for Image Style Classification. MM ’16, pp. 402-406.
[3] S. Matsuo and K. Yanai. CNN-based Style Vector for Style Image Retrieval. ICMR’16, pp. 309-312.

Constructing a Comprehensive Online Platform for Corpus Studies in Taiwan

Howard Chen (National Taiwan Normal University)

Researchers have paid increasing attention to the application of corpora on studies related to Chinese and English literatures, linguistics and teaching. A wide variety of Chinese and English corpora have also been constructed, including written, spoken, web corpora as well as corpora for specific purposes. Although there have been many corpus studies in European countries, there are still few corpus-based studies available in Taiwan [3][4][5]. The lack of corpus-based language studies might be due to the following reasons: (1) the sizes of various corpora are getting bigger, thus it is more challenging for language researchers to process the large data; (2) researchers have difficulties finding available corpora for further research; (3) researchers in general are not very familiar with the PC-based corpus analysis tools, and the PC tools were somewhat limited in their corpus processing capacities; (4) researchers do not have easy access to some useful corpus processing platforms in which they can upload and analyze their own corpora.

Based on the urgent needs of language researchers and students in Taiwan, we examined various corpus analysis tools and discovered that a more robust and comprehensive corpus analysis platform is needed to better help researchers and students in humanities to analyze large language data more efficiently and easily. With the help of the experts in humanities and computer science, this project attempted to build a useful corpus platform to facilitate corpus-based language, literature, and language learning research in Taiwan.

This project developed two corpus analysis platforms. They were based on two well-known corpus analysis engines- NoSketch Engine and CQPweb [2]. Both tools were web-based free tools and are open for downloading. Also, we further collected many available open source corpora (BAWE, BASE, Wikipedia, and so forth) and load them to these two corpus platforms. After the two platforms were built, these corpus tools were made available for various users to conduct corpus research works.

Researchers and students not only can use the various existing Chinese and English corpora but also can apply for their research accounts and upload their own corpora. They can further use the available corpus analysis tools to analyze their data.

To enrich the content of the two corpus platforms, we also applied the method of “web as corpus” to collect more web corpora in order to improve the corpus resources of these platforms [1]. These WAC corpora include various types of written and specialized corpora (Chinese for specific purposes and English for specific purposes).

Moreover, in the future we will further add more functions to our platforms, including keywords analysis, n-gram extraction and semantic analysis. In addition, we will also collect and analyze user behaviors based on the logs of these two corpus platforms, such user information can help us to further improve the services provided in these two sites.

It is expected that the construction of the corpus platforms can be helpful for the researchers and students in studying English literature, linguistics and teaching. We also expect that with the help of these websites, humanity researchers in Taiwan can get access to more comprehensive and powerful corpus analysis platforms to facilitate their data analyses and further enhance their research outcomes.

References

[1]. Baroni, M., & Bernardini, S. (2004, May). BootCaT: Bootstrapping Corpora and Terms from the Web. In LREC.

[2]. Hardie, A. (2012). CQPweb—combining power, flexibility and usability in a corpus analysis tool. *International journal of corpus linguistics*, *17*(3), 380-409.

[3]. McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

[4]. O'keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and*

language teaching. Cambridge University Press.
[5]. Reppen, R. (2010). *Using Corpora in the Language Classroom*. Cambridge University Press.

Counting the Traditional Japanese Musical Scales: Analyzing Folk Songs from Chugoku and Kyushu Districts

Akihiro Kawase (Doshisha University)

Introduction

The main purpose of this study is to estimate the structures of traditional Japanese music in order to embody the concept of Japanese music culture. We apply a scale detection method based on Seiichi Tokawa's scale theory to musical corpora of traditional Japanese folk songs. A scale is a series of musical notes in ascending or descending order, which is an important element for describing the total system (Tonesystem) and capturing the characteristics of the music culture. A fundamental aspect of musical analysis is to clarify the scale in a musical piece. In our past research, we implemented a musical scale detection method based on Tokawa's scale theory, and applied to 1,794 song pieces of Japanese folk songs from Honshu (Kawase and Tokosumi 2011), and compared with Tokawa's classification results from the 251 songs. It turned out that the distributions of the two sets of results almost match. This confirms Tokawa's hypothesis that most Japanese music is the *Yo* type, only a small amount of the music is the *In* type, and the *Ryu-kyu* type exists very rarely, and provided quantitative data to support it (The terms will be described later).

However, analyzes targeting small areas in Japan were not implemented. In this study, we carry out classification experiment of Japanese musical scales for both Chugoku district (the westernmost region of Japan s largest island of Honshu) and Kyushu district (geographically located in the southern part of Japan).

Tokawa's scale theory

Seiichi Tokawa, a Japanese music theorist, classified the Japanese musical scales into four major types: *Yo* (陽), *In* (陰), *Kon-go* (混合), and *Ryu-kyu* (琉球). Basically, each scale has five pitch heights, thus we may generate five scales (or modes) per type by allocating the tonic respectively. **Table 1** is a list of Tokawa's musical scales, which is a transcript from Exploring the Japanese Musical Scale (Tokawa 1990) with partial modifications to pitch height names using the alphabetic letters A-G instead of Japanese characters. The left-hand side of Table 1 presents the scales whose pitch height names are allocated in Tokawa's arrangement, and the right-hand side of Table 1 has the same scales as the left-hand side, but the only difference is that all tonics are lined up to pitch A for comparison purposes. Since Tokawa's theory is not based on the F. Koizumi's tetrachord theory (Koizumi 1958), there is criticism that Tokawa's scale theory cannot capture the characteristics of Japanese music. However, Tokawa's scale theory still provides a versatile system which covers various conventional scale theories under the present circumstances. Accordingly, we constructed a scale detection algorithm based on Tokawa's scale theory (Kawase and Tokosumi 2011), and manipulated it in the Python Environment.

Overview of data

We use the entire musical notes for works included in the "*Nihon Min'yo Taikan*" (Anthology of Japanese Folk Songs, 1944-1993) from each province in the Chugoku district and Kyushu district. In order to digitize the Japanese folk song pieces, we generate a sequence of notes by converting the music score into MusicXML file format. There were 95,154 tones in the sample of 886 songs for Chugoku district and 76,028 tones in the sample of 896 songs for Kyushu district. Therefore, the result of the classification in both districts in this study is that the proportion of *Yo* type is higher than that of Tokawa's result, and the remaining three classes are few.

Table 1: List of Tokawa's musical scales for traditional Japanese music

Yo Type	C mode	C	D	E	G	A	C	A	B	C#	E	F#	A
	D mode	D	E	G	A	C	D	A	B	D	E	G	A
	E mode	E	G	A	C	D	E	A	C	D	F	G	A
	G mode	G	A	C	D	E	G	A	B	D	E	F#	A
	A mode	A	C	D	E	G	A	A	C	D	E	G	A
In Type	E mode	E	F	A	B	C	E	A	Bb	D	E	F	A
	F mode	F	A	B	C	E	F	A	C#	D#	E	G#	A
	A mode	A	B	C	E	F	A	A	B	C	E	F	A
	B mode	B	C	E	F	A	B	A	Bb	D	Eb	G	A
	C mode	C	E	F	A	B	C	A	C#	D	F#	G#	A
Kon-go Type	E mode	E	F	A	B	D	E	A	Bb	D	E	G	A
	F mode	F	A	B	D	E	F	A	C#	D#	F#	G#	A
	A mode	A	B	D	E	F	A	A	B	D	E	F	A
	B mode	B	D	E	F	A	B	A	C	D	Eb	G	A
	D mode	D	E	F	A	B	D	A	B	C	E	F#	A
Ryu-kyu Type	C mode	C	E	F	G	B	C	A	C#	D	E	G#	A
	E mode	E	F	G	B	C	E	A	Bb	C	E	F	A
	F mode	F	G	B	C	E	F	A	B	D#	F	G#	A
	G mode	G	B	C	E	F	G	A	C#	D	F#	G	A
	B mode	B	C	E	F	G	B	A	Bb	D	Eb	F	A

Detection results

If we compile except for song pieces less than 5 pitch heights, it was confirmed that the frequency of *Yo* type is generally large and *Ryu-kyu* type hardly appears. Calculating the percentage for each type, the results of the Kyushu district were 84.94% for the *Yo* type, 13.00% for the *In* type, 1.93% for the *Kon-go* type and 0.13% for the *Ryu-kyu* type. In the Chugoku district, 92.46% for the *Yo* type, 5.84% for the *In* type, 1.70% for the *Kon-go* type and 0.00% for the *Ryu-kyu* type.

Although the material used for the summary is different, according to Tokawa's examination (2007), 251 songs included in *Collection of Japanese Folk Songs* (Iwanami paperback library, 1960) and *Traditional Children's Songs of Japan* (Iwanami paperback library, 1961) were classified as follows: 198 songs (78.88%) for the *Yo* type, 43 songs (17.13%) for the *In* type, 7 songs (2.79%) for the *Kon-go* type, and 3 songs (1.20%) for the *Ryu-kyu* type.

In comparison with various theory of Japanese musical scale theory, for the Chugoku district, the *Ryo* scales were 6.08%, the *Yo* scale ascending order types were 33.70%, the *Yo* scale ascending order types were 16.06%, the *Min'yo* scales were 23.97%, the *In* scale ascending order types were 0.61%, the *In* scale descending order types were 3.77%, and no *Ryu-kyu* scales were detected. On the other hand, for Kyushu district, the *Ryo* scales were 7.08%, the *Yo* scale ascending order types were 23.29%, the *Yo* scale descending order types (in other words the *Ritsu* scales) were 20.21%, the *Min'yo* scales were 23.68%, the *In* scale ascending order types were 0.64%, the *In* scale descending order types (in other words *Miyako-bushi* scales) were 8.37%, and *Ryu-kyu* scales were 0.13%.

In the future study, we will resolve the scale detection method and assemble a database of Japanese folk songs to support musical information, including scale information, historical information, and geographic information, in order to promote folk song research from an engineering perspective.

Acknowledgment

This work was mainly supported by the Japanese Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (15K21601).

References

[1] Kawase, A., and Tokosumi, A. (2011). Scale Detection for Japanese Music Focused on the Pitch Sequence of Musical Scales. In Proc. 6th Spring Conference of JSKE 2011, 21C-07.

[2] Koizumi, F. (1958). *Studies on Traditional Music of Japan 1*, Ongaku no tomo sha.
[3] The Society for Research in Asian Music (1982). *The Series of Research in Asian Music 9*, Tokyo, Ongaku no tomo sha.
[4] Tokawa, S. (1990). *Exploring the Japanese Musical Scale*, Tokyo, Ongaku no tomo sha.

Development of System to Encourage Data Sharing and Usage among Process on Historical Study with Linked Data

Satoru Nakamura (University of Tokyo)

Introduction

Recently, as represented by Open Science or Open Data, the movement to make some data accessible to various stakeholders is growing in necessity. Linked Data, which is a method to allow sharing and usage of data, has played an important role in the movement. There have been many attempts to apply Linked Data to humanities and historical study. CIDOC CRM[1] provides an object-oriented conceptual model that allows researchers and curators to describe and retrieve the information required in museum activities. Historical Event Markup and Linking Project (Heml) [2] provides a format to describe historical events, and software which visualizes those data with timelines and maps. Moreover, Europeana[3] has developed a framework for its metadata which called Europeana Data Model (EDM), and allows access to resources possessed by more than 3,000 cultural facilities. Kiourt et al. [4] developed a system called "DynaMus" that enables virtual exhibition. This system uses data from Europeana such as form and size of resources.

This study aims to the application of this technology into data sharing and usage among process on historical study. We divide historical study into three types of processes. First one is “Management” process, where professionals such as archivists assess, collect, organize, preserve and provide access to historical resources. Second one is “Research” process, where historians aim to study about the past and find new knowledge based on historical resources. Third one is “Exhibition” process, where curators share historical resources and research achievement with the public and community. These processes are common in terms of using historical resources, but each of them requires different expertise and stakeholders. Therefore, these are basically conducted independently. We use Linked Data in order to encourage the efficient coordination among those processes.

Proposed System

Fig. 1 shows an overview of proposed system. This system stores data among processes on historical study integrally on the web, by describing them as Resource Description Framework (RDF). Each historical resource is represented by URI, and data provided from each process is described as metadata of the resource. This study calls data achieved from ”Management” process as “catalog data”, which is basically bibliographic information of historical resources, such as title and created date. Data achieved from ”Research” process is called “research data”, which can be defined and accumulated based on research objectives of each historian, such as research notes and finding. “Exhibition data” represents data used in “Exhibition” process. This represents information such as explanations of resource or additional information related to them such as keywords, in order to help audiences to understand contents or background of resources. In addition, the usage of Web API such as SPARQL, which is a query language of RDF, enables to develop applications for different stakeholders and processes in historical study. This approach of “single source and multiple usage” encourages to share and reuse of data and to enhance efficient coordination of processes in historical study.

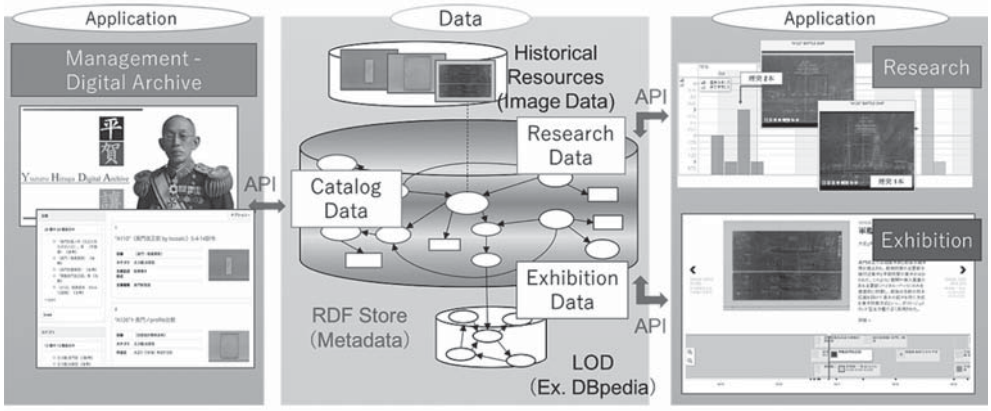


Figure 1: Overview of developed system

Case Study

We applied proposed approach to a case of historical study with “Yuzuru Hiraga Digital Archive”. Yuzuru Hiraga was a vice admiral of Japanese Imperial Navy and a president of the University of Tokyo. He left about 5,500 resources including drawings, engineering documents, reports and others. As for “Management” process in this case, we developed the application which gives access to those resources on the web. This application uses catalog data with standard schema such as Dublin Core. As for “Research” process which uses catalog data from “Management” process, we conducted historical research to confirm the reason of redesign of naval vessels. The method to manage catalog data and research data integrally enabled not only to analyze resources quantitatively with the large amount of catalog data but also to classify them according to the purpose of the researchers. As a result, the reason of difference in the number of funnels between sister vessels was confirmed based on primary resources, such as blueprints and technical reports. As for “Exhibition” process, we developed the application to exhibit those resources on the web by using catalog data and research data. This application gives access to historical resources not only with catalog data but also from several points of view such as timeline, map and network. These achievements demonstrated that proposed approach can encourage to share and reuse of data and to enhance efficient coordination of processes in historical study.

Conclusion

The system to encourage the efficient coordination among processes in historical study with Linked Data was proposed. Developed system allowed users to use data from “Management”, “Research” and “Exhibition” process integrally. This enables to develop applications for different stakeholders and processes. The effectiveness of developed system was demonstrated through the case of historical study with “Yuzuru Hiraga Digital Archive”.

References

[1] CIDOC CRM. “What is the CIDOC CRM?” accessed June 30, 2017. http://www.cidoc-crm.org.
[2] B. Robertson. Exploring Historical RDF with Heml, Changing the Center of Gravity: Transforming Classical Studies Through Cyberinfrastructure, vol 3, no 1, 2009.
[3] A. Isaac, B. Haslhofer. Europeana Linked Open Data --data.europeana.eu, Semantic Web, vol 4, no 3, pp. 291-297, 2013.
[4] C. Kiourt, A. Koutsoudis, G. Pavlidis. DynaMus: A Fully-Dynamic 3D Virtual Museum Framework, Journal of Cultural Heritage, vol 22, pp. 984-991, 2016.

Relationships between Flowers in a Word Embedding Space of Classic Japanese Poetry

Hilofumi Yamamoto (Tokyo Institute of Technology), Bor Hodošček (Osaka University)

Objectives

The present study, which is ongoing work, focuses on exploring the connection between the information content of words and their semantic role within classical poetic Japanese.

Word embedding methods such as Word2Vec (Mikolov et al. 2013) and more recent extensions, including multimodal word distributions in (Chen et al. 2015, Athiwaratkun and Wilson 2017) and richer representations such as doc2vec in (Le and Mikolov 2014), have been shown effective in extracting semantic knowledge that performs well on analogy, semantic similarity, and entailment tasks. In this work, we will quantify the relationship between the information content of a word and its word embedding vector and examine the possibility of using word embedding spaces of classical poetic terms to explain the semantic relationships between terms.

First we will examine if word embeddings encode enough semantic information to be able to determine specific subordinate words via their superordinate concept, represented by a poetic term. For example, flowers frequently appear in classical poetry and tend to be used with specific words. Both ka (aroma) and chiru (fall) are used with flowers such as: “hito wa isa, kokoro mo shirazu, furusato wa, hana zo mukashi no, ka ni nioikeru” by Ki no Tsurayuki (the Kokinshu ¯, No.42; the state of human / hearts I cannot know and yet / the blossoms of this / familiar village still greet / me with the scent of years past / translated by Rodd and Henkenius (1984)). “hisakata no, hikari nodokeki, haru no hi no, shizugokoro naku, hana no chiru ramu” by Ki no Tomonori (the Kokinshu ¯, No.84; the air is still and / sun-warmed on this day of spring– / why then do cherry / blossoms cascade to the earth / with such restless changeful hearts/ translated by Rodd and Henkenius (1984) as well). If ka (aroma) and chiru (fall) are truly relating to specific flowers, it should be possible to acquire the specific name of a flower based on their word embeddings.

Second, we will examine that the residual of A minus a will be nothing if word embeddings are strong enough to extract the specific name a from its superordinate concept A and when $a \in A$ is established.

Methods

We use the Hachidaishū, classical Japanese poem anthologies compiled by the order of the Emperors (ca. 905–1205), comprised of approximately 9,500 poems. Each poem is tokenized by kh (Yamamoto 2007) which divides poem texts into tokens using a classical Japanese dictionary. In order to examine the notable relationships between ‘ka’ (fragrance), ‘chiru’ (fall), we look at the cosine similarity scores between terms in the word embedding space generated by Word2Vec. The Word2Vec model was generated using the Word2Vec implementation in gensim 2.0.0 (R˘eh ˚u˘rek and Sojka 2010). A 50-dimensional skip-gram model with negative sampling was used with context window covering the whole poems.

Results

As a result of measuring the cosine distances between ka (fragrance) and other words, also between chiru (fall) and other words, the list of the former relationship indicates ume (plum) and the latter indicates sakura (cherry) as their corresponding flowers (Table 1 and 2).

As the flower of summer, we could obtain tachibana (the flower of orange), while after removing the vector of tachibana, we could not obtain any names of flowers (Table 3).

Table 1: Top five words similar to ka (fragrance) and the reverse examination of ka by using the term ume (plum). Each value indicates the cosine similarity between each word pair.

	ka (fragrance)		*ume* (plum)	
1	*ume* (plum)	0.96	*ka* (fragrance)	0.96
2	*niofu* (smell)	0.94	*niofu* (smell)	0.92
3	*nushi* (patron)	0.92	*kakine* (fence)	0.90
4	*chirasu* (make fall)	0.90	*nushi* (patron)	0.90
5	*moru* (raise)	0.89	*moru* (raise)	0.90

Discussion

The lists clearly shows that ka (fragrance) is related to ume (Mizutani 1983:130). In terms of chiru (fall), the latter result replicates well-established knowledge in the literature that falling flowers denote sakura (cherry) and not ume (plum), and it is discernable that sakura (cherry) relates to chiru (fall), which indicates that people at the time lamented falling sakura (cherry) (Katagiri 1983: 84). We next looked at whether the Word2Vec word embedding creates a vector space where geometric algebra is possible and vector distances in the space hold certain semantic meaning. Among summer flowers, the tachibana (the flower of orange) is very famous. We expect that if we subtract tachibana out from the summer vectors, the resulting space will be devoid of relationships between natsu (summer) and hana (flower). By calculating with the relational expressions summer + flower and summer + flower − tachibana, the operations conducted by Word2Vec have been shown to reproduce our current understanding of the relationships between flowers and seasons as well as some emotions associated with them. As shown in Table 3, the summer + flower operation indeed includes tachibana and it should be noted that the summer + flower − tachibana operation did not include any remarkable values between summer and flower.

Table 2: Top five similar words to chiru (fall) and the reverse examination of chiru by using the term sakura (cherry). Each value indicates the cosine similarity between each word pair. pn. means 'place name'.

	chiru (fall)		*sakura* (cherry)	
1	*moru* (raise)	0.95	*yamazakura* (mountain cherry)	0.82
2	*sakurabana* (cherry blossoms)	0.94	*Yoshinoyama* (pn.)	0.82
3	*Yoshinoyama* (pn.)	0.93	*chirasu* (make fall)	0.80
4	*yahe* (eight fold)	0.93	*izure* (any)	0.80
5	*yamazakura* (mountain cherry)	0.92	*sakurabana* (cherry blossoms)	0.79

Table 3: Operations relating to natsu (summer), hana (flower), and tachibana (orange).

	flower + summer		flower + summer - orange	
1	*yadosu* (to dwell)	0.90	*yoru* (night)	0.69
2	*kaoru* (to smell)	0.90	*hikari* (light)	0.68
3	*tachibana* (orange)	0.89	*kohori* (ice)	0.67
4	*odoroku* (to surprize)	0.88	*tsura* (face)	0.66
5	*ushirometashi* (to feel guilty)	0.87	*harafu* (pay)	0.66
6	*katsura* (name of tree)	0.87	*fuyu* (winter)	0.66
7	*migaku* (polish)	0.87	*suzushi* (cool)	0.65
8	*issoshi* (more)	0.87	*akeshi* (to dawn)	0.65
9	*haku* (to sweep away)	0.87	*moru* (to leak)	0.65
10	*Musashino* (pn.)	0.86	*niwa* (garden)	0.65

Conclusion

We conducted the experiments using approximately 9,500 classical Japanese poem texts in order to examine the possibilities of extracting subordinate terms from superordinate concepts based on word embeddings. We found that the model also allows us to extract specific subordinate words based on the superordinate concept of classical terms such as: when the distance between two terms such as tachibana (orange) and natsu (summer) is close enough, the superordinate concept A indicates the subordinate concept a. We could therefore verify that it allows us to extract the concrete name from its superordinate concept.

References

[1] Athiwaratkun, Ben and Andrew Gordon Wilson (2017) “Multimodal Word Distributions,” ArXiv e-prints, April.
[2] Chen, Xinchi, Xipeng Qiu, Jingxiang Jiang, and Xuanjing Huang (2015) “Gaussian Mixture Embeddings for Multiple Word Prototypes,” CoRR, Vol. abs/1511.06246, URL: http://arxiv.org/abs/1511.06246.
[3] Katagiri, Yoichi (1983) Utamakura utakotoba jiten (Dictionary of poetic vocabulary), Vol. 35 of Kadokawa shojiten, Tokyo: Kadokawa Shoten.
[4] Le, Quoc V. and Tomas Mikolov (2014) “Distributed Representations of Sentences and Documents,” CoRR, Vol. abs/1405.4053, URL: http://arxiv.org/abs/1405. 4053.
[5] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013) “Efficient Estimation of Word Representations in Vector Space,” CoRR.
[6] Mizutani, Sizuo (1983) Goi (Vocabulary), Vol. 2 of Asakura Nihogo Shin-Kōza, Tokyo, Japan: Asakura Shoten.
[7] R˘eh ˚u˘rek, Radim and Petr Sojka (2010) “Software Framework for Topic Modelling with Large Corpora,” in Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50, Valletta, Malta: ELRA, May, http://is.muni. cz/publication/884893/en.
[8] Rodd, Laurel Rasplica and Mary Catherine Henkenius (1984) Kokinshū - A Collection of Poems Ancient and Modern, Boston MA USA: Cheng and Tsui Company.
[9] Yamamoto, Hilofumi (2007) “Waka no tame no Hinshi tagu zuke shisutemu / POS tagger for Classical Japanese Poems,” Nihongo no Kenkyu / Studies in the Japanese Language, Vol. 3, No. 3, pp. 33–39.

Mapping Dickens's Novels in a Network of Words, Topics, and Texts: Topic Modelling a Corpus of Classic Fiction

Tomoji Tabata (University of Osaka)

Introduction

Topic modelling is a machine learning method for uncovering hidden semantic structures in a corpus of texts. Based on a probabilistic inference algorithm called latent Dirichlet allocation (Blei, et al. 2003), the technique makes it possible to identify sets of frequently co-occurring words, or topics that characterize a text as well as classify texts into groups defined by inferred sets of strongly associated topics. One of the major advantages topic modelling has over traditional key-word detection techniques employed in many stylometric or corpus linguistic studies is that, despite the fact that topic modelling is an unsupervised method for detecting groups of words that tend to co-occur in a set of texts, generated models do not simply help classify texts according to the degree of similarity between texts in terms of topic composition, but also enable us to visualize complex yet meaningful interrelationships between vocabulary items, topics, and more importantly, association between topics and texts in the form of network diagrams or heatmap dendrograms.

Topic Modelling in Digital Humanities

Meeks and Weingart (2012) posit that, "a synecdoche of digital humanities", topic modelling is "distant reading in the most pure sense" since it is focused on a collection of texts in its entirety instead of individual documents. The online *Journal of Digital Humanities*, Volume 2, Number 1, features some exemplar applications of topic modelling in the humanities scholarship, beginning with lucid introductions to topic modelling for the humanities by Blei (2012) and Brett (2012). The journal issue covers topics such as "topic modelling and figurative language" (Rhody, 2012), "his- tory of literary scholarship" (Goldstone and Underwood, 2012), "historical dynamics" (Schmidt, 2012), and a technical review of MALLET, a machine learning language toolkit (Graham and Mil- ligan, 2012). The *JDH* issue was immediately followed by the *Poetics* journal's special number on topic modelling (Vol. 41, 6th Issue). The *Poetics* articles include works on significant themes in 19th-century literature (Jockers and Mimno, 2013), "how democracy handles terrorist threats" (Bonilla and Grimmer, 2012), diachronic changes in language usage (McFarland, et al., 2013), cross-national comparison of disciplinary development (Marshall, 2013), among others. The two journals's coverage illustrates how topic modelling has become an indispensable core of digital humanities research.

Research Method

The proposed study applies topic modelling to a corpus of major 18th Century and 19th Century British fiction (Osaka Reference Corpus for HIstorical/Diachronic Stylistics: ORCHIDS) with a view to analyzing latent semantic structures underlying in the Dickens's fiction and mapping Dickens in the network of words, topics, and texts. What is of special interest is that by means of this approach it is now possible to shed new light on thematic structures composed by a large number of infrequent words, which would otherwise escape the net of key-word statistics due to infrequency of occurrence. Pre-processing treatments of the corpus texts were carried out before running MALLET on the text set: the first stage of pre-processing includes text segmentation (i.e. all texts have been sliced into a consecutive 2,000 word segments with last segments shorter than 2,000 words being discarded for consistency's sake) and removal of all punctuation marks and mark-up tags since the present version of MALLET is not capable of handling them. Frequent proper nouns have been added to the list of stop words so that they do not make it into topics generated from the corpus. The number of topics extracted from the data set is another major factor that would likely affect outputs as well as interpretation of topics. Different settings have been tested so that generated topic models have good enough granularity, in the sense that the models are sensitive enough to not overlook fine shades of

latent semantic patterns in the corpus: the number of topics generated were varied from 15 to as many as 200.

Results and Discussion

Table 1 lists 50 most prominent topics discovered from the corpus with their constituent key words. The complex interrelationships between the topic by way of shared key words can only be visible when the data is projected on to a network diagram as in Figure 1. With regard to the topics that are situated towards the outskirt of the map (3, 0, 32, 6, 4, 35, 43, 49, 46, 40, etc.), their interpretation can be comparatively straightforward. The topics which find themselves in the most densely populated center may defy our scrutiny unless we consult Table 1.

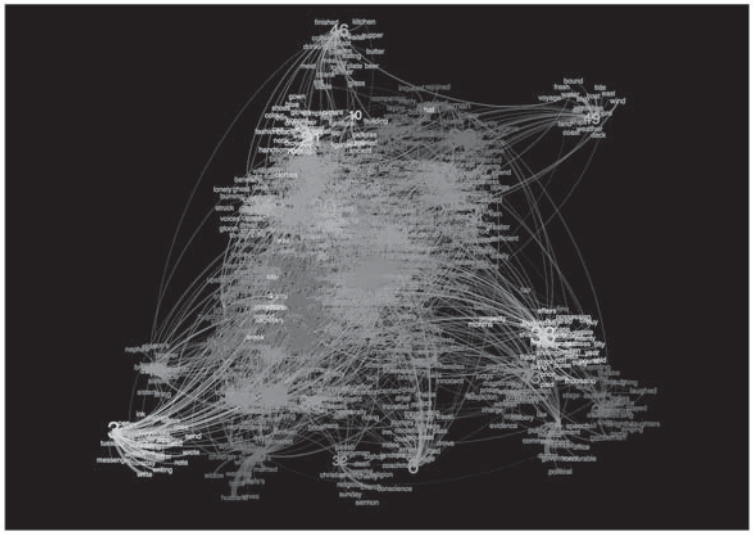


Figure. 1: Network diagram representing interrelationships between topics by way of constituent key words

Figure. 2, a network diagram visualizing association between topics and texts, shows authorial clus- ters as well as diachronic shift in topics in the corpus. The 18th Century texts are located in the left-hand side of the diagram, while the 19th Century texts lean towards the right. Textual cluster- ings emerge in an even more marked pattern: when the topic composition details of each text are compared using a heatmap diagram with hierarchical cluster analysis as in Figure. 3, three distinct clusters appear. The Dickens cluster is clearly differentiated from other 19th Century authors's texts and 18th Century fiction. The 19th Century fiction and the 18th Century texts can be told apart with the exception of Thackeray's *Barry Lyndon* and *Vanity Fair*, which find themselves attached to a small 18th Century cluster consisting of Defoe, Richardson, and Swift. Of special interest with regard to the style of Dickens's novels is that Dickens's texts are characterized by their frequent recourse the following topics: 48) element of suspense, 9) facial/bodily gestures, 16) child, 31) description, 5) motions, and 45) quotative markers.

Table 1: List of 50 topics extracted from the Classic Fiction corpus and their key words

No.	Topic label	Key words, or words with the greatest weight
0	wifehood	wife woman husband married women house life knew people children told widow world day family love like poor way
1	people	people several time part number city houses died dead house called manner persons particular plague town observed certain sick
2	venue	town place general nice way half city hundred thousand war honour body time day part march case mountains number
3	correspondence	letter read write letters written time writing day paper wrote send answer morning hand received pen long friend hope
4	gentlemen's society	doctor men public house office friend day member parliament gentlemen work honourable right government low people question gentleman members
5	motions	head like way returned looked looking asked mind fire hand round right home pretty found coming look pocket taking
6	journey	horse way road coach time journey inn landlord miles ride carriage passed rode fellow found half town turned coachman
7	narrated entities	head boys legs moment short loud dog hands feet suddenly full length nose running eye box arm countenance body
8	perception	like felt life mind feeling sense sort way feel tone looking fact people consciousness possible strong conscious change usual
9	facial/bodily gestrures	hand face looked head hands looking turned look arm held moment stood voice lips sat laid figure raised arms
10	conspicuous	large high picture ancient like beautiful pictures called work figure painted place building long rooms figures furniture noble top
11	sentiment	love heart aunt thought loved words mind world truth happiness moment hope knew true wrong speak thoughts nature bear
12	core values	character find agreeable certain taste particular life public different effect part air means spirit produced conversation family respect exercise
13	adventures	began place company sooner apartment occasion hero gentleman desired received adventurer understand lieutenant finding ordered situation perceived honour satisfaction
14	intimacy	creature heart family cousin friend answer favour friends thought unhappy case wretch find art soul vile relations sex mine
15	character portraits	major day family friend carriage gentleman poor house honour place gave home square passed dinner men course world friends
16	child	child home time dear day heart happy long poor face night life thought tears mother knew kind children like
17	adversity	life death poor god heart time mind heaven die tears grief hope day soul night comfort miserable misery sorrow
18	involvement	time found gave house friend received opportunity account conversation favour behaviour present regard friendship heart manner acquaintance proposal company
19	school(ing)	eye thought like school pupils read long certain found evening mine heart features smile pupil glance full knew books
20	outdoors	day like sun garden walk trees house long green morning air ground stone round pleasant look village high time
21	appearance	like hair black white looked face round dress look brown large long pair wear wore dressed pretty coat blue
22	night	door night bed house morning window open opened heard hour stairs day time long candle looked shut hours light
23	dramatic present	looks takes look like friend way head hand returns venus time makes right gentleman looking court riderhood sits eye
24	womanly care	poor girl like thought told woman house way pretty stay look care bear gave mistress suppose call creature leave
25	narrated manner	looked speak words moment door voice turned asked hand face spoke heard look thought rose word answer chair sat
26	narrative time	time house place way mind question found case certain words asked view happened interest moment chance morning answered mine
27	omniscient narration	course knew thought told house cousin way word rate doubt men friend asked suppose moment truth felt understand found
28	UNINTERPRETABLE	like bit work mother thought lad time home way look words mind poor right pretty folks long deal heart
29	affect	time felt day home long happy morning half feel feelings place spirits evening longer happiness wished visit hope friends
30	narrated speech acts	answered cries gentleman honour madam began cried company truth friend occasion fellow kind sooner part received account poor convinced
31	description	state general fact manner family society present life business observed subject occasion position course nature highly point establishment character
32	church	church bishop men clergyman god religion poor hospital parish read christian sunday years charity friends dean religious true priest
33	politeness	dear madam hope happy honour pleased love dearest worthy kind favour heart mother god occasion goodness mind time servant
34	abstract values	power means passion virtue present nature degree human honour purpose easily greatest design certain youth proper short sufficient difficulty
35	socialization	ladies play gentlemen company music people dance gentleman conversation evening girls fine pretty table laugh part played dancing sing
36	feminine discourse	dear mind hope way returned like time word find understand beg head suppose pray look glad case manner question
37	journaling	day dined morning secretary told home letter dine queen night town business like court hope send answer late ten
38	monetary	money pounds hundred business thousand years year paid sum day time life twenty live ten fifty account worth price
39	learnedness	world men nature book knowledge like life generally learned years mind learning history true education books age natural human
40	familial	father mother brother uncle son daughter father's family years child world time mother's way hand right nephew children story
41	gentleman	gentleman replied gentlemen inquired morning hat door exclaimed friend friends large countenance looked round fat blue dear course time
42	topography	street people streets place night men like town crowd houses windows city way iron window black places yard dirty
43	court of law	justice law time case court evidence prison found murder brought knew guilty told committed judge crime life hands trial
44	cognition	thought friend general subject family possible like attention deal perfectly sort manners obliged replied agreeable object understand suppose moment
45	quotative markers	cried replied returned time gentleman friend looking head hand face rejoined door word turning look round manner fellow way
46	dining	dinner table wine tea glass eat bread water cold fire company breakfast supper hot kitchen day drank sat drinking
47	exploits	men found told way place began thought knew resolved gave time rest island lay work brought several part called
48	elements of suspense	light fire dead dark strange cold sound long like stood darkness heard deep round wind voice fell strong feet
49	voyage	sea ship water boat time wind shore ships land lay found tide voyage night weather men vessel sail day

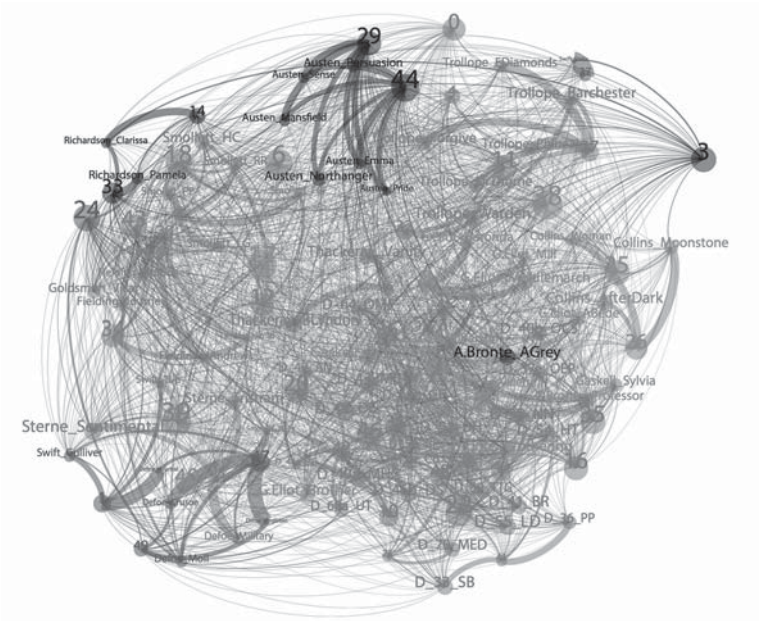


Figure. 2: Network diagram representing interrelationships between topics by way of constituent key words

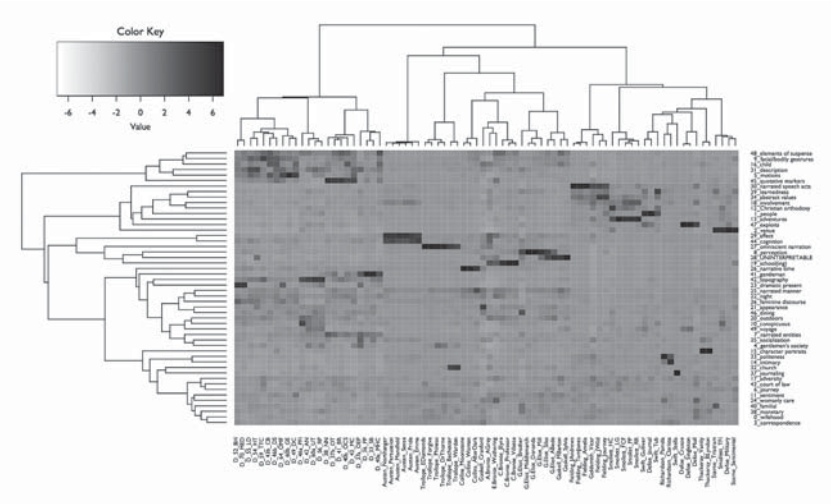


Figure. 3: Heatmap of 50 topics across 78 texts (with mean weights scaled)

Emerging results from this research are expected to open up a new avenue of inquiry into key semantic patterns in the ORCHIDS, thereby suggesting a possibility of building a bridge between findings from machine learning text mining and linguistic stylistics, between distant reading and close reading, leading to an empirical interplay of insights that can benefit textual analysis.

References

[1] Blei,D.M.,Ng,A.andM.Jordan2003.LatentDirichletAllocation.*JournalofMachineLearningResearch*,3: 993–1022.

[2] Blei,D.M.(2012a)Topicmodelinganddigitalhumanities,*JournalofDigitalHumanities*,2(1),2012. Available at http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/

[3] Blei,D.M.(2012b)ProbabilisticTopicModels.*CommunicationsoftheACM*,55(4):77–84. doi:10.1145/2133806.2133826

[4] Bonilla, T. and J. Grimmer (2013) Elevated threat levels and decreased expectations: How democracy handles terrorist threats, *Poetics*, 41 (6): 650–669.

[5] Brett,M.R.(2012)TopicModeling:ABasicIntroduction,*JournalofDigitalHumanities*,2(1). Available at http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/
[6] Crymble, A. (2012) Review of Paper Machines, produced by Chris Johnson-Roberson and Jo Guldi, *Journal of Digital Humanities*, 2 (1). Available at http://journalofdigitalhumanities.org/2-1/review-papermachines-by-adam- crymble/
[7] Deerwester, S., Dumais, S., Landauer, T., Furnas, G., and R. Harshman (1990) Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41 (6): 391–407.
[8] DiMaggio, P., Nag, M., and D. Blei (2013) Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts, *Poetics*, 41 (6): 570–606.
[9] Fothergill, R., Cook, P., and T. Baldwin (2016) Evaluating a Topic Modelling Approach to Measuring Corpus Similarity, *Procedings of LREC 2016, Tenth International Conference on Language Resources and Evaluation*, 23th-28th May, 2016, Portorož, Slovenia. Available at http://www.lrec- conf.org/proceedings/lrec2016/pdf/154_Paper.pdf
[10] Goldstone, A. and T. Underwood (2012) What Can Topic Models of PMLA Teach Us About the History of Literary Scholarship? *Journal of Digital Humanities*, 2 (1). Available at http://journalofdigitalhumanities.org/2- 1/what-can-topic-models-of-pmla-teach-us-by-ted-underwood-and-andrew-goldstone/
[11] Graham, S. and I. Milligan (2012) Review of MALLET, produced by Andrew Kachites McCallum, *Journal of Digital Humanities*, 2 (1). Available at http://journalofdigitalhumanities.org/2-1/review-mallet-by-ian-milligan- and-shawn-graham/
[12] Hofmann,T.(1999)ProbabilisticLatentSemanticAnalysis,*UAI'99ProceedingsoftheFifteenthConferenceon Uncertainty in Artificial Tntelligence*, 289–296.
[13] Jaworska,S.andA.Nanda(2016)DoingWellbyTalkingGood:ATopicModelling-AssistedDiscourseStudyof Corporate Social Responsibility, *Applied Linguistics*, 2016: 1–28. doi: 10.1093/applin/amw014
[14] Jockers, M. L. and D. Mimno (2013) Significant themes in 19th-century literatureOriginal Research Article, *Poetics*, 41 (6): 750–769.
[15] McFarland, D. A., Ramage, D., Chuang, J., Heer, J., Manning, C. D., and D. Jurafsky (2013) Differentiating language usage through topic models, *Poetics*, 41 (6): 607–625.
[16] Manning, C. D. and H. Schütze (1999) *Foundations of Statistical Natural Language Processing*. Cambridge, MA.: The MIT Press.
[17] Marshall. E. A. (2013) Defining population problems: Using topic models for cross-national comparison of dis- ciplinary development, *Poetics*, 41 (6): 701–724.
[18] Meeks,E.andS.B.Weingart(2012)TheDigitalHumanitiesContributiontoTopicModeling,*Journalof Digital Humanities*, 2 (1). Available at http://journalofdigitalhumanities.org/2-1/dh-contribution-to-topic-modeling/
[19] Miller, I. M. (2013) Rebellion, crime and violence in Qing China, 1722 - 1911: A topic modeling approach, *Poetics*, 41 (6): 626–649.
[20] Mohr, John W. and Petko Bogdanov (2013) Introduction—Topic models: What they are and why they matter, Poetics, 41 (2013), 545—569. http://dx.doi.org/10.1016/j.poetic.2013.10.001
[21] Mohr, J. W., Wagner-Pacifici, R., Breiger, R. L., and P. Bogdanov (2013) Graphing the grammar of motives in National Security Strategies: Cultural interpretation, automated text analysis and the drama of global politics, *Poetics*, 41 (6): 670–700.
[22] Murakami,A.,Thompson,P.,Hunston,S.,andD.Vajn(2017forthcoming) Whatisthiscorpusabout?,*Corpora*, Issue 12.2. Pre-publication copy. Available at http://pure-oai.bham.ac.uk/ws/files/25529512/TopicModelPaper_ACCEPTED_VERSION.pdf
[23] Papadimitriou, C., Raghavan, P., Tamaki, H., and S. Vempala (1998) Latent Semantic Indexing: A probabilistic analysis, *PODS '98 Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, 159–168. doi:10.1145/275487.275505
[24] Rhody,L.M.(2012)TopicModelingandFigurativeLanguage,*JournalofDigitalHumanities*,2(1). Available at http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-

language-by-lisa-m-rhody/
[25] Salton, G. and M. McGill (eds.) (1983) *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.
[26] Schmidt, B. M. (2012) Words Alone: Dismantling Topic Models in the Humanities, *Journal of Digital Humani- ties*, 2 (1). Available at http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/
[27] Tangherlini, T. R. and P. Leonard (2013) Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research Original, *Poetics*, 41 (6): 725–749.
[28] Törnberg, A. and P. Törnberg (2016) Combining CDA and topic modeling: Analyzing discursive connections between Islamophobia and anti-feminism on an online forum, *Discourse & Society 2016*, Vol. 27 (4): 401–422.

Analyzing Features for the Detection of Happy Endings in German Novels

Fotis Jannidis, Isabella Reger, Albin Zehe, Martin Becker, Lena Hettinger, Andreas Hotho (University of Würzburg)

Abstract

With regard to a computational representation of literary plot, this paper looks at the use of sentiment analysis for happy ending detection in German novels. Its focus lies on the investigation of previously proposed sentiment features in order to gain insight about the relevance of specific features on the one hand and the implications of their performance on the other hand. Therefore, we study various partitionings of novels, considering the highly variable concept of "ending". We also show that our approach, even though still rather simple, can potentially lead to substantial findings relevant to literary studies.

Introduction

Plot is fundamental for the structure of literary works. Methods for the computational representation of plot or special plot elements would therefore be a great achievement for digital literary studies. This paper looks at one such element: happy endings. We employ sentiment analysis for the detection of happy endings, but focus on a qualitative analysis of specific features and their performance in order to gain deeper insight into the automatic classification. In addition, we show how the applied method can be used for subsequent research questions, yielding interesting results with regard to publishing periods of the novels.

Related Work

One of the first works was on folkloristic tales, done by Mark Finlayson, who created an algorithm capable of detecting events and higher-level abstractions, such as villainy or reward (Finlayson 2012). Reiter et al., again on tales, identify events, their participants and order and use machine learning methods to find structural similarities across texts (Reiter 2013, Reiter et al. 2014).

Recently, a significant amount of attention has been paid to sentiment analysis, when Matthew Jockers proposed emotional arousal as a new “method for detecting plot” (Jockers 2014). He described his idea to split novels into segments and use those to form plot trajectories (Jockers 2015). Despite general acceptance of the idea to employ sentiment analysis, his use of the Fourier Transformation to smooth the resulting plot curves was criticized (Swafford 2015, Schmidt 2015).

Among other features, Micha Elsner (Elsner 2015) builds plot representations of romantic novels, again by using sentiment trajectories. He also links such trajectories with specific characters and looks at character co-occurrences. To evaluate his approach, he distinguishes real novels from artificially reordered surrogates with considerable success, showing that his methods indeed capture certain aspects of plot structure.

In previous work, we used sentiment features to detect happy endings as a major plot element in German novels, reaching an F1-score of 73% (Zehe et al. 2016).

Corpus and Resources

Our dataset consists of 212 novels in German language mostly from the 19th century[1]. Each novel has been manually annotated as either having a happy ending (50%) or not (50%). The

[1] Source: https://textgrid.de/digitale-bibliothek

relevant information has been obtained from summaries of the Kindler Literary Lexikon Online[2] and Wikipedia. If no summary was available, the corresponding parts of the novel have been read by the annotators.

Sentiment analysis requires a resource which lists sentiment values that human readers typically associate with certain words or phrases in a text. This paper relies on the NRC Sentiment Lexicon (Mohammad and Turney 2013), which is available in an automatically translated German version[3]. A notable feature of this lexicon is that besides specifying binary values (0 or 1) for negative and positive connotations (2 features) it also categorizes words into 8 basic emotions (anger, fear, disgust, surprise, joy, anticipation, trust and sadness), see Table 1 for an example. We add another value (the polarity) by subtracting the negative from the positive value (e.g. a word with a positive value of 0 and a negative value of 1 has a polarity value of -1). The polarity serves as an overall sentiment score, which results in 11 features.

Table 1: Example entries from the NRC Sentiment Lexicon

Word/Dimension	verabscheuen (to detest)	bewundernswert (admirable)	Zufall (coincidence)
Positive	0	1	0
Negative	1	0	0
Polarity	-1	1	0
Anger	1	0	0
Anticipation	0	0	0
Disgust	1	0	0
Fear	1	0	0
Joy	0	1	0
Sadness	0	0	0
Surprise	0	0	1
Trust	0	1	0

Experiments

The goal of this paper is to investigate features that have been used for the detection of happy endings in novels in order to gain insight about the relevance of specific feature sets on the one hand and the implications of their performance on the other hand. To that end, we adopt the features and methods presented in Zehe et al. (2016). The parameters of the linear SVM and the partitioning into 75 segments are also adopted from this paper.

Features. Since reliable chapter annotations were not available, each novel has been split into 75 equally sized blocks, called *segments*. For each lemmatized word, we look up the 11 sentiment values (including polarity, see above). Then, for each segment, we calculate the respective averages, resulting in 11 scores per segment. We group those 11 scores into one feature set.

Qualitative Feature Analysis. As our corpus consists of an equal number of novels with and without happy ending, the random baseline as well the majority vote baseline amount to 50% classification accuracy. Since we assumed that the relevant information for identifying happy endings can be found at the end of a novel, we first used the sentiment scores of the final segment ($f_{d,n}$) as the

only feature set, reaching an F1-score of 67%. Following the intuition that not only the last segment by itself, but also its relation to the rest of the novel are meaningful for the classification, we introduced the notion of *sections*: the last segment of a novel constitutes the *final section*, whereas the remaining segments belong to the *main section*. Averages were also

[2] www.kll-online.de

[3] http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

calculated for the sections by taking the mean of each feature over all segments in the section. To further emphasize the relation between these sections, we added the differences between the sentiment scores of the final section and the average sentiment scores over all segments in the main section. However, this change did not influence the results. This led us to believe that our notion of an "ending" was not accurate enough, as the number of segments for each novel and therefore the boundaries of the final segment have been chosen rather arbitrarily. To approach this issue, we varied the partitioning into main and final section so that the final section can contain more than just the last segment.

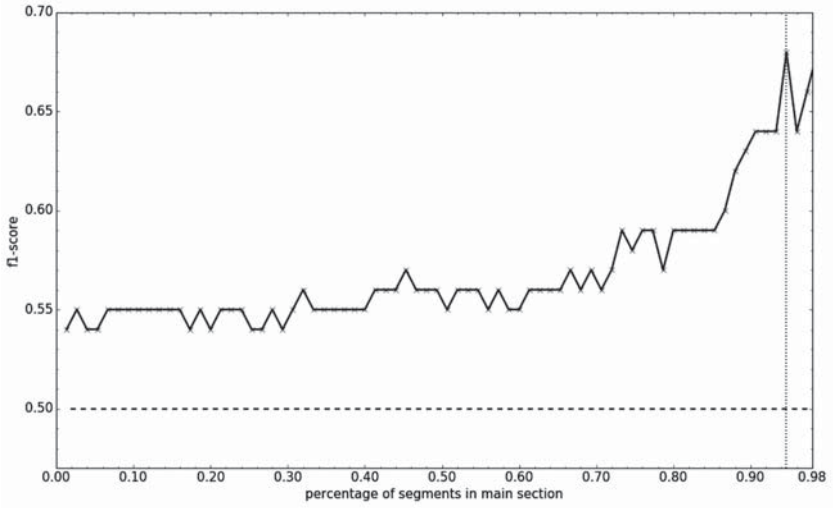


Figure 1 : Classification F1-score for different partitionings into main and final section. The dashed line represents a random baseline, the dotted line shows where the maximum F1-score is reached.

Figure 1 shows that classification accuracy improves when at least 75% of the segments are in the main section and reaches a peak at about 95% (this means 4 segments in the final section and 71 segments in the main section, for a total of 75 segments). With this partitioning strategy, we improve the F1-score to 68% using only the feature set for the final section ($f_{d,\ final}$) and reach an F1-score of 69% when also including the differences to the average sentiment scores of the main section ($f_{d,\ main\ -\ final}$).

Since adding the relation between the main section and the final section improved our results in the previous setting, we tried to model the development of the sentiments towards the end of the novel in a more profound way. For example, a catastrophic event might happen shortly before the end of a novel and finally be resolved in a happy ending. To capture this intuition, we introduced one more section, namely the *late-main section,* which focuses on the segments right *before* the final section, and used the difference between the feature sets for the late-main and the final section as an additional feature set ($f_{d,\ late\ -\ final}$).

Using those three feature sets, the classification of happy endings reaches an F1-score of 70% and increases to 73% when including the feature set for the final segment.

Table 2: Classification F1-score for the different feature sets

Features	Results
1) Final segment feature set	67%
2) Final segment feature set and difference to main section	67%
3) Final section feature set with final section of length 4	68%
4) Feature set 3 and difference to main section	69%
5) Feature set 4 and difference between late-main section and final section	70%
6) Feature set 5 and final segment feature set	73%

Table 2 summarizes these results and shows that the addition of each feature set leads to small improvements, amounting up to a F1-score of 73%. While we saw that the classification performs best when the final section consists of 4 segments, we also observed that quite a few novels could be correctly classified with several different partitionings. On the other hand, some novels could not be predicted correctly with any choice of partitioning. An example is *Twenty Thousand Leagues Under the Sea* by Jules Verne which evidently has a happy ending with clearly identifiable boundaries, but an extremely short one, consisting only of about 250 words. These observations show that the notion of a novel's "ending" is highly variable and can differ considerably from text to text.

Correlation with Publication Dates. This raises the question whether we can use the sensitivity of our approach to this kind of variability in order to better understand the characteristics of the novels in our corpus. As an example, we studied whether different section partitionings are in any way correlated with the publication date of a novel. In order to keep the results as interpretable as possible, we focused on one single feature set: the sentiment scores of the final section. In a first attempt, we divided our corpus into four subgroups, distinguishing novels published before 1830 (65 novels), between 1831 and 1848 (31 novels), between 1849 and 1870 (29 novels) and after 1871 (87 novels). This split resulted in similarly sized portions and did not yield a strong bias towards happy/unhappy endings in any period.
Figure 2 shows that the best classification is again obtained when about 95 - 98% of the segments are in the main section, regardless of the time period. Therefore, the best section split

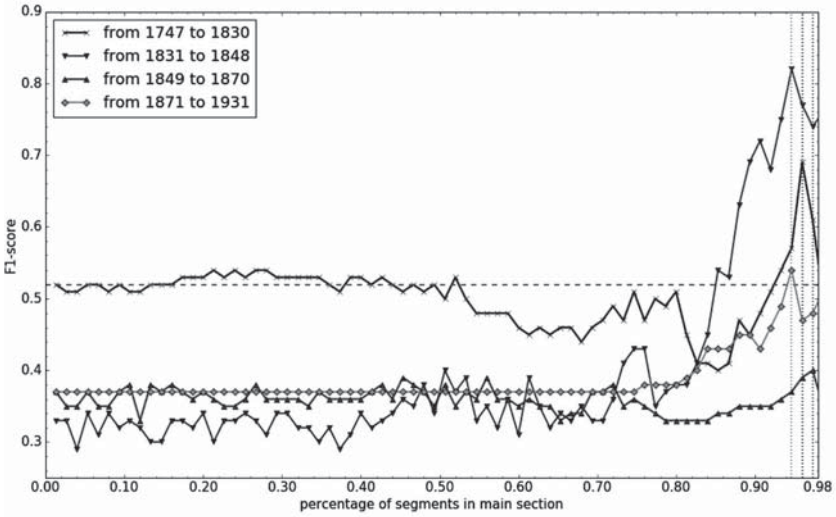


Figure 2 : Classification F1-score for different partitionings into main and final section. Each line denotes novels from a different time period. The dashed line represents the random baseline for the time period starting from 1871. Random baselines for the other periods yield slightly worse results and are omitted. The dotted lines show where the maximum F1-score is reached for the respective time periods.

point is not correlated with the publication date of a novel. What is striking, however, is the fact that the novels published after 1848 yield considerably lower scores than the novels published before that year, mostly even below the baseline. This indicates a correlation between publication date and automatic classification quality, i.e. novels published before the period of Realism are more easily classifiable in terms of having a happy ending than realistic novels. A possible explanation is that many novels of that earlier period are more schematically structured.

We are aware that the number of novels for each of the time spans is rather small, so that those findings can only be regarded as exploratory insights. Nevertheless, these preliminary results show that the automatic detection of happy endings, even with only one rather simple feature set, can uncover dependencies to other properties of novels that are highly interesting for literary studies.

Conclusion and Future Work

The automatic detection of happy endings as a major plot element of novels is a valuable step towards a comprehensive computational representation of literary plot. Our experiments show that different features based on sentiment analysis can predict happy endings in novels with varying but reasonable quality. Even though our approach is still rather simple, we showed that it can potentially lead to substantial insights for literary scholars.
Future work may cover improving our classification by accounting for the high variability of endings in novels and may also include further leveraging our approach to study the characteristics of different novel collections in-depth.

References

[1] Elsner, Micha (2015): “Abstract Representations of Plot Structure”, in: *Linguistic Issues in Language Technology* 12 (5).
[2] Finlayson, Mark A. (2012): *Learning Narrative Structure from Annotated Folktales*. PhD thesis, Massachusetts Institute of Technology.
[3] Jockers, Matthew L. (2014): “A novel method for detecting plot”. http://www.matthewjockers.net/2014/06/05/a-novel-method-for-detecting-plot/ [Access date 25. August 2016].
[4] Jockers, Matthew L. (2015): “The rest of the story”. http://www.matthewjockers.net/2015/02/25/the-rest-of-the-story/ [Access date 25. August 2016].
[5] Mohammad, Saif / Turney, Peter (2013): “Crowdsourcing a Word-Emotion Association Lexicon”, in: *Computational Intelligence* 29 (3): 436-465.
[6] Reiter, Nils (2013): *Discovering Structural Similarities in Narrative Texts using Event Alignment Algorithms*. PhD thesis, Heidelberg University.
[7] Reiter, Nils / Frank, Anette / Hellwig, Oliver (2014): “An NLP-based Cross-Document Approach to Narrative Structure Discovery”, in: *Literary and Linguistic Computing* 29 (4): 583–605. 10.1093/llc/fqu055.
[8] Schmidt, Benjamin M. (2015): “Commodius vici of recirculation: the real problem with Syuzhet”. http://benschmidt.org/2015/04/03/commodius-vici-of-recirculation-the-real-problem-with-syuz het/ [Access date 25. August 2016].
[9] Swafford, Annie (2015): “Problems with the Syuzhet Package”. https://annieswafford.wordpress.com/2015/03/02/syuzhet/ [Access date 25. August 2016].
[10] Zehe, Albin / Becker, Martin / Hettinger, Lena / Hotho, Andreas / Reger, Isabella / Jannidis, Fotis (2016): "Prediction of Happy Endings in German Novels", in: *Proceedings of the Workshop on Interactions between Data Mining and Natural Language Processing 2016*.

Machine-Learning Approaches to Literary Works: Novels of Sir Arthur Conan Doyle

Ayaka Kuroda (University of Osaka)

Sir Arthur Conan Doyle is well-known as the author of Sherlock Holmes series, and a number of studies have been carried out to examine characters' personalities or investigatory details in the Holmes series. However, his strong inclination for historical fiction has not widely been recognized. To clarify his works' features, it is important for us to focus not only on Holmes series but also on novels of other genres. Also, we have to pay attention to stylistic aspects of his novels and short stories. In the humanities, quantitative investigation using computers is attracting attention, because it is believed to be able to complement existing literary studies from a new point of view. The work by Burrows (1987), who investigates the works of Jane Austen using Principal Components Analysis (PCA), was one of the pioneering studies which obtain great benefit from literary computing. Such a stylometric approach falls within a practice of 'distant reading' (Morretti, 2000), which is an antonym of 'close reading'.

This study attempts to find differences between Doyle's historical fiction and detective fiction through statistical analysis of words in texts with special reference to word frequency pat- terns, and provides a new perspective for literary scholarship. Two types of machine-learning analyses are conducted to highlight linguistic or stylistic features that distinguish the two text genres.

First, we used Random Forests to find genre-specific 'key words', which are significantly more frequent in one category than in another. In corpus linguistics, several studies have pro- posed various measures for identifying key words in the form of significance testing, such as log-likelihood ratio and chi-square value. These methods, however, can be inadequate when the corpus/corpora under investigation include(s) long novels, because proper nouns or words related to particular themes/settings tend to get an excessively high keyness score. To deal with this problem, Tabata (2015) proposed a method that uses Random Forests, an ensemble learning algorithm for classification or regression, to extract words contributing to classification. In this study, all texts were automatically classified into two groups according to word frequency patterns with an accuracy of 96.58%.

Table 1: Random Forests result summary

Call:			
randomForest(formula = text.group ~ .,			
data = tbl, proximity = T, importance = T, ntree = 100000)			
Type of random forest:			classification
Number of trees:			100000
No. of variables tried at each split:			22
OOB estimate of error rate:			3.42%
Confusion matrix:			
	HF	SH	class.error
HF	97	0	0.0000000
SH	5	44	0.1020408

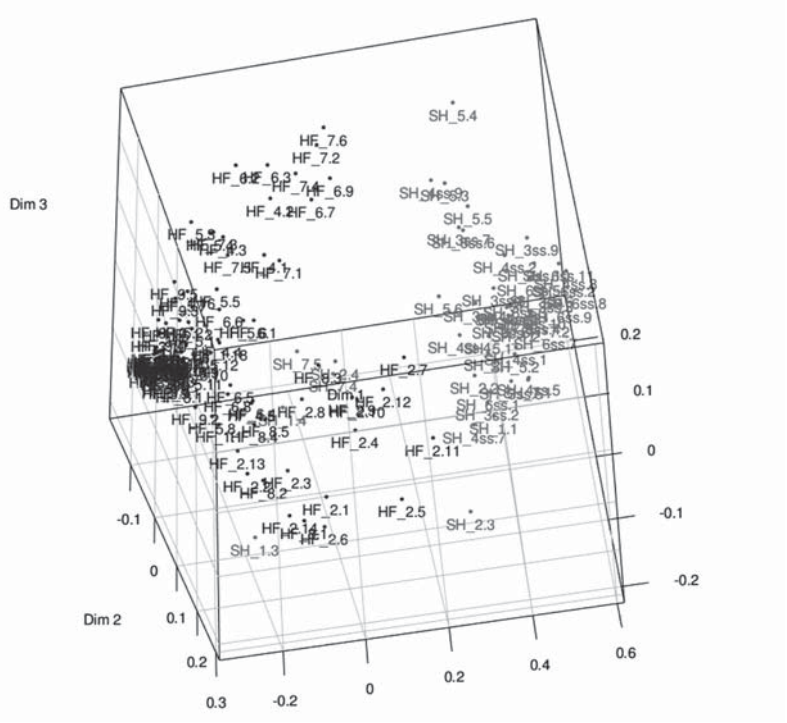


Figure 1: Relationships between topics and their constituent words (50 topics extracted)

The above Figure 1 is a 3D scatter plot which is generated based on Multi Dimensional Scaling (MDS). According to this figure and Table 1, small number of the Sherlock Holmes text segments (five out of forty-nine 10,000-word blocks) were misclassified as historical fiction. We examined these segments to account for the reason why they were mislabeled as historical fiction. Two out of the five misclassified blocks belong to Part II of *A Study in Scarlet*, the part recounted in the style of a 'retrospective narrative'. Part II of the novel explains past events and the background of the murder, while Part I describes Holmes's investigation and his great success. The same can be said of other three segments: those sections also accounts for the causes of the case in 'retrospective narrative'. We therefore conclude that these retrospective narrative has similar stylistic features to historical fiction. To make these features more specific, we made a list of words which contributed to the classification. Based on the list, we found some key words of historical fiction: 'cried' and body-part words such as 'head', 'faces', 'arms' and 'eyes'. Key words of detective fiction, on the other hand, include 'case', 'found', or words related to housing and furniture such as 'house', 'door' and 'chair'.

Second, MALLET was used in conjunction to build topic models based on Latent Dirichlet allocation (LDA) proposed by Blei, Ng and Jordan (2003). Topic modeling is one of text mining tools using machine-learning algorithm for discovering 'latent topics' of texts, a cluster of co-occurring words. By intuition, readers infer the topic of a text by its word-usage pattern: if a text includes words 'game', 'players', 'ball' and 'goal', readers presume that the text is about 'sports'. Also, we can unconsciously understand that one text may contain two or more topics, such as 'sports and medicine' or 'sports and economy'. Topic modeling simulates this intuition by mathematical calculation of what words constitute a topic and to what extent a text exhibits a particular topic(s). That is to say, Random Forests could find some 'keywords', but topic modeling can reveal not only keywords but also relationships between them.

The granularity of topic models depends on the number of topics extracted: with a very small number of topics, generated models may end up being too coarse to interpret, while with a very large number of topics, resulting models will likely include insufficient number of items, too few words to observe their complicated relationships from. The ideal setting is depending on the target corpora's scale or the purpose of modeling, so there is no optimal number.

In this study, we experimented with two different settings: we compared results obtained from 20 topics and 50 topics. Network graphs were drawn to show complex relationships be- tween topics, those between topics and their constituent words, and association between topics and texts. We found some topics which emerge more frequently in detective fictions. The 'Housing and Furnitures' topic was among the most important features of detective fiction, a result that is in keeping with findings obtained with Random Forests. A topic labeled as 'Characters of Holmes Series and Criminal Investigation' was also one of the key topics of detective fiction. It includes 'Holmes', 'Watson', 'case', 'inspector' and 'police', but character names and common nouns are mixed in one topic when 20 topics were extracted. When we built 50 topics, however, they were divided into two topics and became independent of each other, so the co-occurrence of each topics became clearer.

We used two algorithms, Random Forests and topic modeling, to apply quantitative analysis to Conan Doyle's historical fiction and detective fiction. What emerges from our investigation are linguistic features that differentiate between these two text genres. Since the linguistics features of his novels have hardly been discussed in the existing literature, we hope to be able to provide a new perspective for literary investigation into the novels of Sir Arthur Conan Doyle.

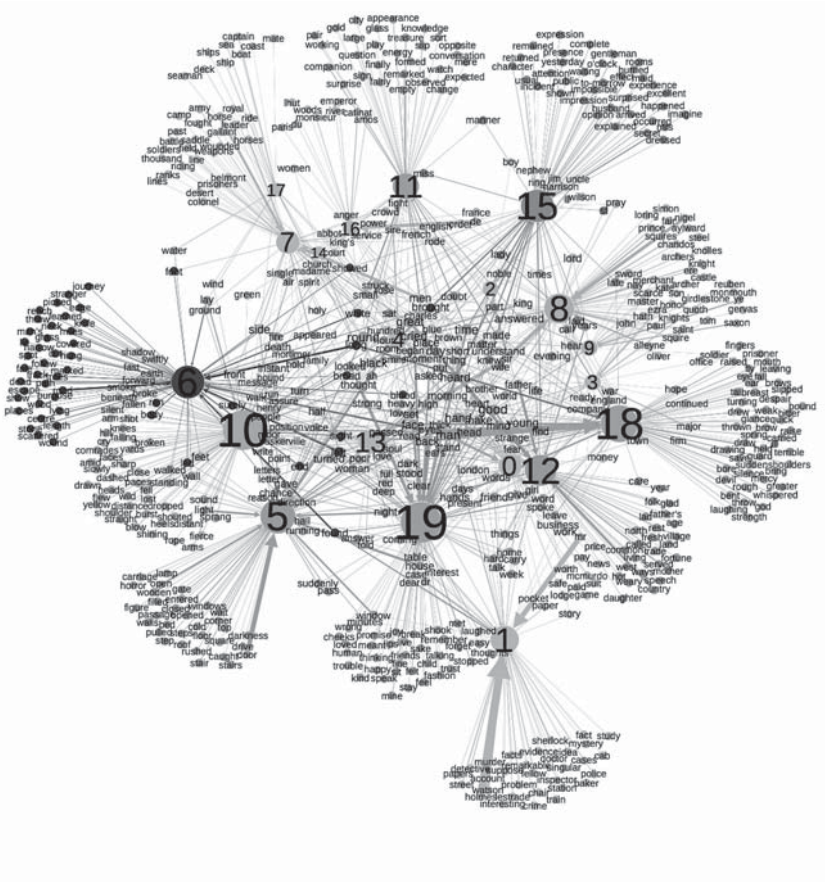


Figure 2: Relationships between topics and their constituent words (20 topics extracted)

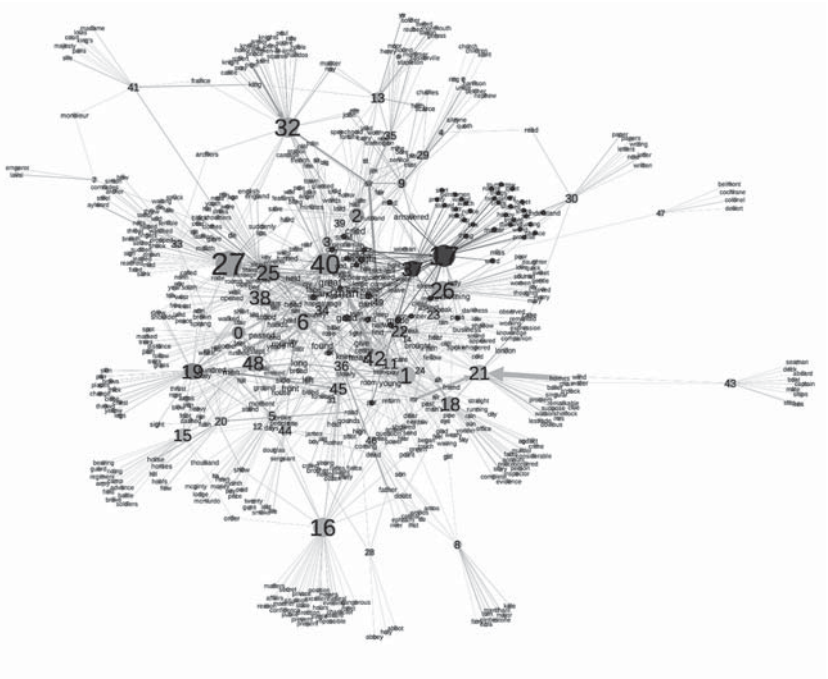


Figure 3: Relationships between topics and their constituent words (50 topics extracted)

Human-assisted OCR of Japanese Books with Multistep Microtasks

**Kosetsu Ikeda (Chiba University),
Kiyonori Nagasaki (International Institute for Digital Humanities),
Atsuyuki Morishima (University of Tsukuba)**

In Recent years, digital transcription for Japanese books with crowdsourcing become popular in Japan. Because the Japanese language contains thousands of characters some of which are similar to each other, digital transcription needs much human power even if the transcribing objects are printing with type. There are two common approaches to transcribe. One is the using online collaboration tools for specific manuscripts, articles, books, and so on. The other is the human-assisted optical character recognition (human-assisted OCR) approach, where crowd workers correct the results of OCR software or directly transcribe characters that the OCR fails to transcribe. The human-assisted OCR approach is compatible with micro-volunteering, that is, to divide a big task into small ones to be able to be solved by crowd workers. It will enable to transcribe in parallel and minimize individual burden.

This paper reports our project for transcribing the book collections of National Diet Library (NDL) in Human-assisted OCR approach with Crowd4U, a micro-volunteering and crowdsourcing platform.

Our project is unique in three ways. First, we divide digital transcription into multistep microtasks to reduce a requestor's burden and improve output quality. In the first step, crowd workers assess a part of the object as to whether it should transcribe. This step try to exclude disturbance like figures, tables, and so on from following steps. In the second step, crowd workers judge whether some OCR results are correct or not at once. Because to correct OCR results for every single character is inefficient, crowd workers make a rough judge in this step. This approach is effective if OCR boast high precision. In the last step, crowd workers correct OCR results as every single character. In all steps, we adopt using majority votes to absorb some worker's mistake.

Second, we try to extend the human-assisted OCR approach by distributing microtasks in many ways other than just showing tasks on PC screens. Most of existing projects ask crowd workers to perform such tasks on the specific Web page. In contrast, we try to extend the approach by distributing tasks in many other ways. For example, we distribute tasks to the task-on-the-floor (shortly TOF) system, where people walking perform tasks while walking on the floor, and to the smartphone lockscreen application, where people who want to unlock their smartphones perform tasks. These are very useful for simple tasks, like the first step and the third step tasks. It helps to keep the number of performed microtasks stable.

Third, we deal with Japanese books which have thousands of characters. This raises an interesting question because people performing microtasks may not pay a great attention to the task compared to the case where they perform tasks on the specific Web page.

In Crowd4U, if the workers logging in, their names will appear in the list of contributors on our web site. Also, they can leave themselves anonymous.

The main contributions are as follows. First, we explain our ongoing project whose approach is novel. To the best of our knowledge, our project is the first to try to transcribe with multistep microtasks and distribute tasks in many ways other than PC screens in the digital library domains. Second, we show our preliminary results that suggest that our approach is effective to some extent although the Japanese language contains thousands of characters some of which are similar to each other.

Creating Geotagged Humanities Data via Mobile Phone: Opportunities and Challenges

David Joseph Wrisley (New York University Abu Dhabi) Mario Hawat, Dalal Rahme (American University of Beirut)

Discussions of crowdsourcing in the digital humanities often concern textual data, in particular, transcription of manuscript data. Theorists point to the ability of "groups to out-perform individual experts, [and] outsiders [to] bring fresh insights to internal problems" (Brabham). In terms of data collection, mobile devices are described as freeing humans to collect significantly larger samples of data in space, in particular in the domains of citizen participation or infrastructure management. Crowdsourced data collection via mobile devices might seem like an ideal match, and yet the technique presents numerous challenges, foregrounding the necessity for digital humanities research to be vigilant about the changing modes of cultural knowledge production in "complex computational societies" (Berry/Fagerjord). Our paper discusses three ways that collecting data for digital humanities projects via mobile phones introduces new levels of data complexity: (1) the tension between ontological precision and on the fly human-intelligence tasking (2) the paradoxes of urban mapping research: redundancy, coverage and privacy (3) the impact of human behavior with socially embodied devices in sensitive environments.

Our paper discusses the experience of the "Linguistic Landscapes of Beirut" project (llbeirut.org). Our data consist of geotagged photos captured via mobile phones. The 2000+ images of urban multilingual writing were collected in two phases over two semesters in 2015- 16 year by a team of some thirty undergraduate researchers. They include metadata that are both automatically generated (time, latitude, longitude, image-size and phone model) and user-generated (language, script and linguistic features). A third post-processing phase of the project began in late 2016 focusing on more granular annotation and the transcription of the multilingual text found in the photos in YAML format.

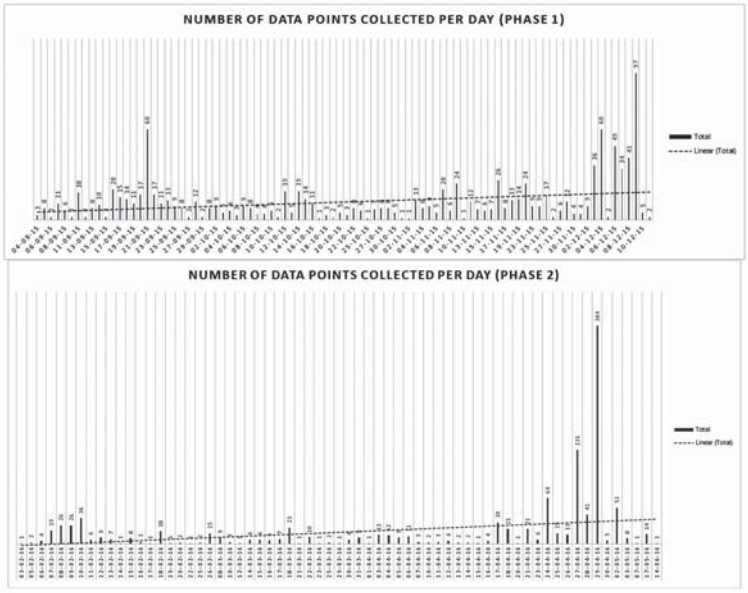
Our project corresponds to two aspects of Brabham's taxonomy of crowd participation: knowledge discovery and distributed human-intelligence tasking. Compared to other mobile application-based linguistic landscaping, our project is smaller in scale, yet richer in metadata (Lingscape). Whereas our data resemble volunteered geographical information (VGI), in that data collectors were free to capture images anywhere within an urban perimeter, it might be better described as a semi-directed, collaborative mapping since choice on the app data form was constrained to a bounded set of fields dictated by specific research questions.

First, an issue often discussed with VGI is the resultant data quality. Indeed, in the post-processing phase some inconsistencies in the classification of the samples were uncovered, but a more salient issue in the social creation of data was what we might call an ontological "shift," whereby the crowd avoided some categories and moved to nuance them in open comment fields. This was partially solved by iterative analysis of the data and collective reassessment of data fields. By keeping the number of "on the go" human-intelligence tasks to a minimum in the data entry form, we believed to be assuring data quality, but there were still a number of unclassifiable examples. We have attempted to deal with ambiguity in the post-processing phase via image annotation in order to qualify and identify these samples.

Second, whereas it has been argued that the geographical information from historical sources is inherently ambiguous and reflects user bias (Dossin et al.), the geolocation in our project was an automatic feature of the form builder. The data collection was left intentionally unstructured--participants were not obliged to use a specific sampling method. On the one hand, this meant that data accumulated along routes of the team's daily mobility. Again, iterative analysis of the data during the collection process showed the zones of greatest data density. Seeing this visualized in real time in-app and in web mapping environments encouraged some to venture

out into uncharted spaces in the city. On this point, we would argue that user specificity of the data reflected less a bias than contributing to the overall diversity of neighborhood coverage. This, however, required the anonymization of the dataset, since participants tended to collect data around their places of work and residence.

Third, discussions of VGI often mention user motivation. We did notice that although some "super users" (Causer and Wallace) emerged in the initial phases of the data collection, this abated, perhaps due to the social pressure not to over-perform in the pedagogical setting. Although pacing the data collection out over time was encouraged, the numbers of image samples captured tended to intensify at the end of each academic term as shown in Figures 1 and 2. We do not believe this to be a problem for the linguistic landscape data, but could have an impact on other projects that are more time sensitive.



Figures 1 and 2: A bar chart representing the number of data points collected per day over the two phases of data collection, with spikes at the end of each semester.

The mobile application method of the data collection "on the go" did allow us to scale up the data rather quickly, but it also revealed a cultural discomfort with photography in a divided city with highly visible security mechanisms. On the one hand, we realized that many of the photos were being taken out of car windows, a phenomenon for which the automatically-generated GPS_SPEED field provided some insight (Hawat). A more astonishing crowd pattern emerged however when we examined the aggregate of the data against the visible security mechanisms in municipal Beirut and found an uncanny avoidance of almost all of the secured zones of the city.

Figure 3: A map showing the aggregate of the "Linguistic Landscapes of Beirut" data (red) along with a rough depiction of the visible security mechanisms (yellow) following Harb et al.

This leads us to question what other blind spots exist in human behavior with mobile devices and encourages us to rethink the assumption of full freedom of movement through urban space when it comes to data collection.

Works Cited (all links accessed on 30 June 2017)

[1] Berry, D. and Fagerjord, A. (2017). *Digital Humanities: Knowledge and Critique in a Digital Age*, Polity.

[2] Brabham, D. (2013). *Crowdsourcing*, MIT Press.

[3] Causer, T. and Wallace, V. (2012). Building A Volunteer Community: Results and Findings from Transcribe Bentham. *DHQ*, vol. 6.2 http://www.digitalhumanities.org/dhq/vol/6/2/000125/000125.html

[4] Dossin, C., N. Ningning Kong and B. Joyeux-Prunel. (2015). Applying VGI to collaborative research in the humanities: the case of ARTL@S, *Cartography and Geographic Information Science*. http://dx.doi.org/10.1080/15230406.2016.1216804

[5] Gilles, P. And Purschke, C. (2017). Lingscape. Mobile Application. https://lingscape.uni.lu/

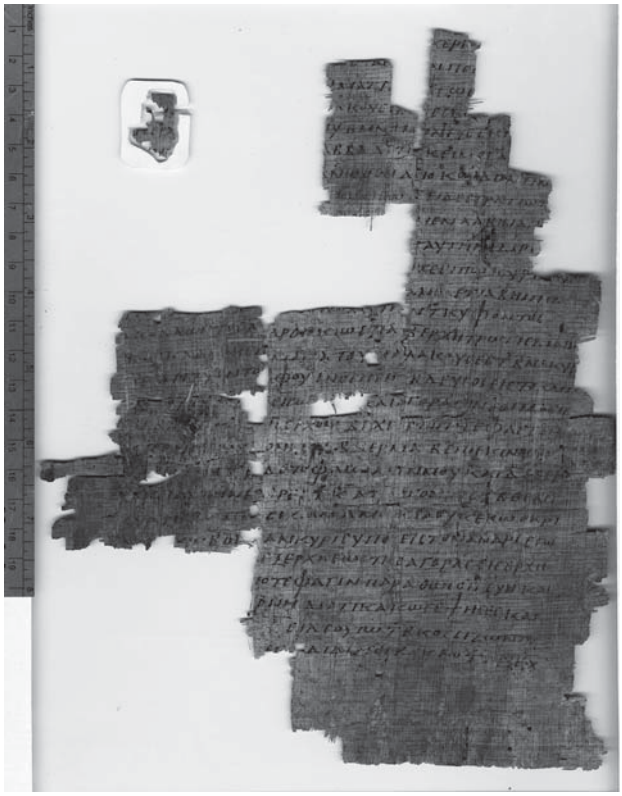
[6] Hawat, M. (2016). "Mapping Mobility." Blog. http://www.mariohawat.com/blog/2016/5/16/moving- magic?p

[7] Harb, M., Gharbieh, A. and Fawaz, M., eds. (2009). *Beirut / Mapping Security*. IABR.

[8] March, J. (2016). "On Choosing a Mobile Platform in the Digital Humanities." Blog. https://digitalfellows.commons.gc.cuny.edu/2016/02/29/on-choosing-a-mobile-platform-in-the-digital-humanities/

Another Kind of Mime. Another Kind of Digital Humanities

James Brusuelas (University of Oxford)



As papyrologists, we are part of a long-standing initiative to integrate our discipline into the digital sphere through websites, algorithmic textual analysis, and general applications based on visual and textual data, an initiative that goes back to the 1990's. The goal has been to produce tools that have impact on the research of professional scholars and students. The field of papyrology has been quite successful in this respect. However, digitally creative and artistic output aimed at communities beyond academia is under represented in the evolving Digital Humanities. And so, when the Arts and Humanities Research Council in the UK announced a unique funding challenge to produce creative digital content, we proposed *Broken Scenes: Resurrecting Ancient Fragmented Voices Through Animation*. For Classics, only small scale attempts at integrating animation have been conducted, such as the Panoply project (www.panoply.org.uk), which animates static scenes on ancient vases for enriching the teaching and learning about well known Greek myths. There has been no merging of classics scholars and professional artists from the animation industry to bring about a digital transformation of ancient culture that has cultural impact beyond the academic community. Can the artistry of animation be integrated with papyrology to visualize and help resurrect lost ancient performative voices preserved only in fragmented papyri?

In 2014 the *Oxyrhynchus Papyri* published P. Oxy. 5189, a 6th century CE Greek Mime. A popular form of entertainment in antiquity, the Mime was a type of sketch comedy performed on stage or in the street and known for its vulgar humour and improvisation. Despite such popularity, the textual evidence of 'low' literature and entertainment of the Graeco-Roman world did not make it into the familiar channels of the mediaeval transmission, a process in which selected authors were copied by scribes and scholars. Our knowledge of popular entertainment is thus largely dependent on papyri. Reconstruction of this fragmentary evidence, however, is a challenge due to the literal holes and gaps in the text, which we supplement as best we can, according to our knowledge of the Greek language and literature. In this case, there is only a narrative description

of the movements and actions of the characters, including the actors' lines, with the following discernible content: 1) clothing (perhaps cross-dressing); 2) eating and (badly) cooking; and 3) beating and slapping, perhaps triggered by bad cooking. As a reading text, it is rather awful due to its fragmented state. But this was never a text to be read. It was meant to be acted. It was meant to be seen. Accordingly, the idea of animation as a visual, rather then purely textual, means of reconstruction emerged.

The purpose of this paper is to introduce the animated short film *Trashy Humour: A Comedy in Pieces*. In collaboration with Acme Filmworks, a BAFTA and Academy Award winning animation studio in Los Angeles, CA, papyrology at the University of Oxford engaged in a different kind of Digital Humanities. Based on the remaining content of P.Oxy. 5189, we experienced the full production process, from script and storyboards to multiple digital animatics and final edits. Over the course of production we thus addressed a variety of research questions. From both digitally artistic and academically philological perspectives, what does it mean to digitally transform a papyrus fragment, its fragmented content, into an animated short film? Who is the audience? Is it a kind of documentary? Or is it entertainment and thus subject to poetic license and adaptation? More importantly, why do it all? And when finished, what does one do with it? Can it be used for teaching? In then end, what place does animation have in Digital Humanities?

Matrix and Graph Operations for Relationship Inference: An Illustration with the Kinship Inference in the China Biographical Database

Chao-Lin Liu (Harvard University/ National Chengchi University), Hongsu Wang (Harvard University)

Biographical databases contain diverse information about individuals. Person names, birth information, career, friends, family and special achievements are some possible items in the record for an individual. The relationships between individuals, such as kinship and friendship, provide invaluable insights about hidden communities which are not directly recorded in databases. We show that some simple matrix and graph-based operations[1] are effective for inferring relationships among individuals, and illustrate the main ideas with the China Biographical Database[2] (CBDB).

Relationship Matrices

Assume that we have n different individuals and m different relationships. Let P_i, $i \in \Pi=\{1,\cdots,n\}$, represent the individuals, and R_j, $j \in \Lambda=\{1,\cdots,m\}$, represent the relationships. We can store the relationships between the individuals with a matrix $\boldsymbol{T}$, in which a cell $T_{1,2}$ is the relationship between P_1 and P_2 that is recorded in a database. In CBDB, an R_1 can be **basic**, e.g., "F" which represents a "father-of" relationship, and an R_2 can be **extended**, e.g., "DHB" which represents a "daughter's husband's brother"[3] relationship while the participating individuals in between are unknown. In the matrix in Figure 1, we have $T_{1,2}$= "F", and that means P_2 is a father of P_1. That $T_{2,3}$= "DHB" means P_3 is P_2's daughter's husband's brother.

$$\mathbf{T} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} & \mathrm{F} & \\ & & \mathrm{DHB} \\ & & \end{bmatrix} \end{array}$$

Figure 1: A relationship matrix for kinship

A **relationship matrix** like that in Figure 1 does not have to be symmetric. In addition, a cell in the matrix may need to accommodate multiple values when necessary. When we convert the relationship matrix directly from a biographical database, the original matrix may be asymmetric, because the information stored in databases are not symmetric. If we wish, we can build a symmetric matrix from a raw matrix like that in Figure 1, by checking and combining $T_{i1,i2}$ and $T_{i2,i1}$ for any combination of $i1 \in$ "Π" and $i2 \in$ "Π". In this process, we might find conflicting or even parallel relationships between two individuals, which will need domain experts for further inspection. A person may have parallel (or multiple) relationships with another due to many reasons.

[1] Chao-Lin Liu, Tun-Wen Pai, Chun-Tien Chang, and Chang-Ming Hsieh (2001) Path-planning algorithms for public transportation systems, *Proc. of the Fourth Int'l IEEE Conf. on Intelligent Transportation Systems*, 1061-1066.

[2] China Biographical Database: https://projects.iq.harvard.edu/cbdb

[3] Here, D stands for "daughter", H for "husband", and B for "brother".

$$\mathbf{M} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \end{array} \qquad \mathbf{M}^2 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{array}$$

Figure 2: A binary relationship matrix and its square

We can define operations for relationship inferences with a relationship matrix (RM). The simplest choice is to convert cells that have known relationships into 1s and others into 0s. We convert the matrix in Figure 1 into the **M** and compute its square in Figure 2. Note that the 1 in $\mathbf{M}^2$ is informative. It indicates that P_3 is P_1's father's daughter's husband's brother.

The fact that the powers of an RM are informative is a natural result of the definition of multiplication of matrices. Let $\mathbf{M}^2_{x,y}$ denote the cell (x,y) of $\mathbf{M}^2$. We compute$\mathbf{M}^2_{x,y}$ based on the definition in (1). When we convert an RM into a binary RM, a cell is one only if the cell represents an existing relationship. Hence, definition (1) essentially states that P_x and P_y has relationships if there exists a person P_z such that (1) P_z is a relative of P_x and (2) P_y is a relative of P_z. Furthermore, the value of $\mathbf{M}^2_{x,y}$ is the total number of such two-step relationships.

$$\mathbf{M}^2_{x,y} = \sum_{z \in \Pi} \mathbf{M}_{x,z} \times \mathbf{M}_{z,y} \tag{1}$$

In addition to counting the indirect relationships when computing (1), we can concatenate the relationships of $\mathbf{M}_{x,z}$ and $\mathbf{M}_{z,y}$ to obtain a two-step relationship for P_x and P_y. For instance, concatenating "F", "2", and "DHB" to obtain "**1_F_2_DHB_3**", in the computation for the $\mathbf{M}^2$ in Figure 2, will record one indirect relationship between P_1 and P_3 and the intermediate participants. This operation of concatenation and recording can be achieved by software easily.

Applications and Extensions

We can easily generalize the implication of (1) to consider the semantics of the values in the n-th power of a binary RM, **M**, to achieve the following theorem.

Theorem 1. Let **M** denote a binary relationship matrix as we explained above. A person P_x and a person P_y are relatives if there is a $\rho \geq 1$ such that $\mathbf{M}^{\rho}_{x,y} > 0$.

There are 360,000 individuals in the 2015 version of CBDB, and we should be able to identify clusters of individuals that represent families. Although Theorem 1 is applicable for this task, a more efficient method is to apply the ideas of identifying connected components in mathematical graphs. The basic idea is quite simple: individuals that have relationships belong to a family and members of different families cannot have any relationships.

Corollary 1 allows us to find the shortest relationships between two individuals. If the kin relationships in the original RM are genuinely basic, e.g., father, month, son, daughter, husband, wife, brother, and sister, then we could find the most direct relationships between individuals with the corollary.

Corollary 1. Let **M** denote the binary RM of an RM. By finding the smallest σ such that $\mathbf{M}^{\sigma}_{x,y}$ is positive, we identify the relationship between P_x and P_y that has the fewest number of intermediate participants.

In practice, not all kin relationships recorded in historical documents are basic, and extended relationships such as "DHB" may be recorded. As a consequence, a "shortest path" as defined in Corollary 1 may not be the most direct path between two individuals.

We can define the "**length**s" of relationships to quantify the distances between individuals.[4] Take

[4] Similar quantification methods were adopted by Professor Michael Fuller when he designed a backend service for CBDB. Similar quantification methods were also described in the following working paper.

Ke Deng, Peter K. Bol, Stuart M. Shieber, and Jun S. Liu. Building a Kinship Network for Political

"DHB" as an example. That P_y is a "DHB" of P_x means that P_y is one generation after P_x, and we can define the **g-len** of "DHB" as -1[5]. Analogously, the g-len of "F" is 1. The g-len of a path is the sum of the g-lens of all relationships in the path, so the g-len of "F_2_DHB" is 0. The length of a relationship can encode the "sideway" distances as well. We can define the **s-len**s of "DHB" and "F" to be 2 and 0, respectively. In practice, there are only few hundreds of relationships in CBDB, so it is feasible to annotate the relationships with their lengths.

Building Family Networks

The upper part of Figure 3 shows three actual paths[6] between individuals in CBDB. We can construct a family network that is **consistent** with the paths, and provide the network to domain experts for verification.

NAME3_B_NAME4_S_NAME5 (i.e., NAME3's brother is NAME4;
NAME4's son is NAME5) NAME3_FF_NAME1_SS_NAME4_S_NAME5
NAME3_F_NAME2_S_NAME4_S_NAME5

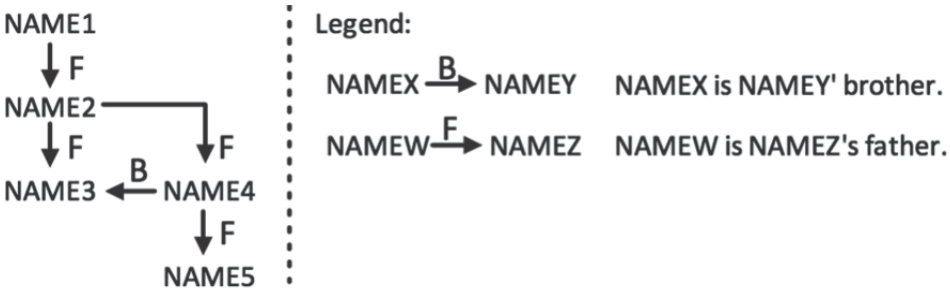


Figure 3: Creating a network with paths

Figures in History.

[5] D is a relationship from an earlier generation to a later generation (-1). Both H and B are relations for the same generations (+0). Hence the g-cost of "DHB" is -1+0+0=-1.

[6] The actual names in Chinese are NAME1 for 呂弸中 (lu3 peng2 zhong1), NAME2 for 呂大器 (lu3 da4 qi4), NAME3 for 呂祖謙 (lu3 zu3 qian1), NAME4 for 呂祖儉 (lu3 zu3 jian3), and NAME5 for 呂喬年 (lu3 qiao2 nian3).

Semi-automatic reconstruction of category for integrated ordinance database

Takanori Kawashima (National Diet Library),
Takashi Harada (Doshisha University)

Introduction

Our project is building "Ordinance Web Archive Database"[1], which collect ordinance of 1727 Japanese local governments and provide full-text search for about one million ordinances. In this database, facet filtering with categories is not implemented although each ordinance has headings like "social welfare" or "agriculture and forestry". Category facet seems to be useful to navigate through complicated law system especially for novice, but it is not provided for the following reason. Because Japanese law does not strictly regulate legal system of local government' ordinance, each municipality is creating their own headings/categories for ordinances. So one municipality has category named "social welfare", and the other has "public welfare", but they contain many similar ordinances.
Actually, the title or content of each ordinance is similar in many municipalities. Kakuta calculated editorial distance for each text of ordinances and found there exists about 500 patterns of ordinances in 802 municipalities [2].

Method

There are about 700 variations in top level headings, and about 5,000 variations in second level headings in character strings, so it is hard to create categories manually. Therefore, we've tried creating it automatically.
In this case, creating categories anew is not necessary, because there are existing categories. What we have to do is reconstructing them and which can be done by clustering them. The outline of the clustering algorithm is as follows.

1. Create feature vector for existing categories
Each ordinance has top level heading (h1) and second level heading (h2). Conversely, each h1 or h2 has several child ordinances. We use title words (noun) in these ordinances as feature, because same kind of ordinance has similar title across municipality. To create ordinance feature vector, CBoW model vector [3] trained with word2vect and Japanese Wikipedia [4] is utilized to absorb the blurring of terms used and reduce dimension of feature vector. The feature vector is sum of CBoW model vectors from words included in child ordinances.

2. Exclude generic categories
Some municipality create less-meaningful category, such as "ordinance collection". To exclude them, all ordinances are redistributed to categories which are most close in Pearson Coefficient of feature vectors. With this, some categories will have no child ordinances and they are excluded.

3. Clustering
We chose k-means++ as clustering method using Pearson Coefficient as distance measure. The numbers of clusters are decided as 10 for h1, and 30 for h2 for test data, as a result of elbow method with try and error. The purpose of this algorithm is to create comprehensive categories, so some arbitrariness is no problem in this case.

4. Decide hierarchy
Average feature vectors are calculated for each category clusters. Then, the parent h1 category clusters of h2 category clusters are decided which has minimum Pearson Coefficient of feature vectors.

5. Name category cluster
The name of a category cluster is decided by name of child categories. Words appearing in more than one-third of child categories are elected as name of categories.
To evaluate the algorithm, ordinances of one prefecture is selected as test data, which consist of 762 h1 headings, 3,058 h2 headings and total of 47,636 ordinances.

Results

An example of generated categories is as shown below.

h1	h2
General Rule	City Regime, Commendation, The Part*
Fire fighting	Fire fighting
Salary	Travel expenses/Remuneration, Personnel affairs, Status of civil servant/headcount, Group of civil servant, Service of civil servant
Finance	Society, Contract, Environment*, House*, Accounting, Medicine for late elderly people*, Tax
Parliament	Election committee
Sentence #	City Regime*
Education	Documents*, Equity Committee*, Agriculture Committee/ Board of Education, Society*, Public announcement ceremony*, Affiliation*, Fixed asset appraisal review committee/Audit committee*, Mayor*, Cultural property, Public relations*
Public enterprise/Construction	Sewage, Public announcement ceremony, City

Categories marked # are h1 category clusters which is not suitable for top level category. The name "Sentence" is derived from "Above Sentence". These categories should be merged to "General Rule", as "City Regime" h2 categories appear twice. Other top level categories are suitable, but "Education" has lot of inappropriate * marked h2 category clusters, which suggest there is need of h1 categories like "committee" or "social welfare".

To evaluate accuracy of h2 category clusters, 489 (1%) randomly selected ordinances are checked whether they are a) under the correct category, b) under irrelevant category, c) can't be decided. The result is a) 48%, b) 43%, c) 7% (conducted by 3 experimenters and category which get 2 or more votes are counted. When the vote broke, it is counted as "c".).

About 60% of ordinances decided as (b) or (c) is under "Finance-Society" or "Education-Mayor". Former category contains many ordinances about children, handicapped, or elderly welfare, which includes financial aspect of policy (child medical expenses payment ordinance), and facility (General welfare center ordinance) or abstract one (Home childcare ordinance). Judged as (b) in experiments includes facility or abstract kind of welfare ordinances, so "Finance" h1 is misleading in this case. Latter category includes variety of ordinances such as "Ordinance to establish a special group for city promotion strategy". It seems "Mayor" category includes ordinances which are directly under mayor's control, or something cannot be categorized to other categories, and this leads the inaccuracy.

Excluding (b) and (c) of these 2 categories, 69% of ordinances are under correct category.

Conclusions and Future Works

Although the accuracy of the result is not high, the method is almost automatic and easily expandable to whole municipality. To improved accuracy, it is possible to redistribute ordinances again to constructed category clusters. For further research, we will ask evaluation of generated

categories by law professional, and it is also useful to modify generated categories by professional. Furthermore, by improving this algorithm, it is possible to apply this to integrated database of other kind.

References

[1] http://joreislis.doshisha.ac.jp/
[2] Kakuta T. Report about the progress and future of eLen project, Journal of Law and Politics, v.252, 2013, p.247-234
[3] Mikolov T, Chen K, Corrado G, Dean J, Efficient Estimation of Word Representations in Vector Space, arXiv:1301.3781, 2013
[4] http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/

Employing Syntactic Features in Authorship Attribution of Three Writers: Alice Bradley Sheldon, Ernest Hemingway, and Theodore Sturgeon

Miki Kimura (Meiji University)

Abstract

This study performs a quantitative authorship attribution analysis on the works of Alice Bradley Sheldon (1915-1987), an American writer of feminist science fiction who primarily used the pen name James Tiptree, Jr., to disguise her true gender. Given that Sheldon masqueraded as the male James Tiptree, Jr., for almost a decade, many critics have discussed the author's gender. Silverberg (1975) insisted, with reference to the style of Ernest Hemingway, that James Tiptree's stories were written by a man. In this study, we performed a quantitative stylistic analysis of Sheldon's work, comparing it with that of Ernest Hemingway and Theodore Sturgeon, a male science fiction writer whose career overlapped with Sheldon's. By comparing all of Alice Sheldon's works with those of Sturgeon and Hemingway, we hoped to find clues about Sheldon's allegedly masculine writing style. The analysis used syntactic variables as discriminants, specifically the distribution of parts of speech (POS). Four kinds of statistical analyses were conducted by two unsupervised algorithms (a cluster analysis and a principal component analysis) and two supervised learning algorithms (using support vector machines and random forests). The study examined inter-author variations between the three designated authors. The results show two kinds of supervised learning methods can detect inter-author variations. Furthermore, based on the multi-dimensional scaling (MDS) plot of the random forests results, Alice Sheldon's style is similar to that of Ernest Hemingway, just as many literary critics aver.

Introduction

This paper presents a case study of a quantitative authorship attribution dealing with the works of Alice Bradley Sheldon (1915-1987), an American writer of feminist science fiction. When she first began writing in 1967, Sheldon used the male pen name James Tiptree, Jr., both to conceal her identity and as a commercial strategy. In this way, she successfully disguised her gender.

During the decade that the name James Tiptree, Jr., concealed Sheldon's identity, many critics discussed the author's gender. The best-known literary critique was Robert Silverberg's introduction to the 1975 collection of Tiptree's stories, *Warm Worlds and Otherwise.* In the 3,000-word essay, titled "Who Is Tiptree? What Is He?", Silverberg (1975) wrote:

Hemingway was a deeper and trickier writer than he pretended to be; so too with Tiptree, who conceals behind an aw-shucks artlessness an astonishing skill for shaping scenes and misdirecting readers into unexpected abysses of experience. And there is, too, that prevailing masculinity about both of them.

Later, after Alice Sheldon's true identity was revealed, Lefanu (1989) also compared Tiptree's stories to Hemingway's, noting that Tiptree's manner of writing was masculine.

In order to investigate the similarities and dissimilarities between Alice Sheldon's works and those of Ernest Hemingway. However, there were many confounding factors when comparing their works, such as different genres and periods. To overcome these problems, a corpus of Theodore Sturgeon's short stories was also compiled. Sturgeon (1918-1985) was an American writer of short science fiction stories for over 30 years and a contemporary of Sheldon's.

Note that if one of the approaches, such as cluster analysis, is sensitive enough to make intra-author discriminations between works attributed to Tiptree and those attributed to Raccoona Sheldon, then the same approach will also detect inter-author variations between Alice Sheldon and the other two authors. On the other hand, it is possible that the author discriminators applied to these corpora differentiate between Alice Sheldon and the other two authors, but fail to discriminate between Alice Sheldon's two pseudonyms. The latter result would mean that Alice Sheldon failed to disguise her true identity by using the pen names Thus, by comparing Alice Sheldon's works under her two pen names with the works of Hemingway and Sturgeon, we

hoped to find clues about Sheldon's allegedly masculine writing style.

Data and Methods

The three corpora compiled for this study contained all of Alice Sheldon's published work under both her pen names (72 works with 865,802 word tokens), all of Hemingway's work (69 works with 271,475 word tokens), and all of Sturgeon's work (222 works with 1,952,895 word tokens). In order to maintain consistent sample sizes, 70 of Sturgeon's works were randomly selected from his corpus.

The emphasis in this research is primarily on variations between the three authors' works. In addition, the study also compared the results of the quantitative stylometry with the opinions of literary critics such as Silverberg (1975), Lefanu (1989) and others. Following the approaches of Hirst and Feiguina (2007), Hou and Jiang (2016), and Kimura (2017), our quantitative analysis used syntactic variables as effective discriminants (specifically, the distribution of parts of speech: POS). We chose unigram of POS as variables in this analysis. Furthermore we carried out two kinds of unsupervised statistical analyses, namely cluster analysis and principal component analysis (PCA), and two kinds of supervised statistical methods, namely support vector machines (SVM) and random forests. The results of these statistical analyses were compared with each other.

Results

Figure 1 shows the result of cluster analysis with multi-scale bootstrap resampling. As revealed by the dendrogram for the cluster analysis, no James Tiptree cluster or Raccoona Sheldon cluster was detected. Alice Sheldon tried to vary her writing style for her two pseudonyms, but the syntactic features used in these analyses revealed no significant variation between the stories published under the two different pen names.

For the unsupervised methods, signs of inter-author variations were also considered, but no visible clusters or groups were seen for Alice Sheldon, Hemingway or Sturgeon. In conclusion, the unsupervised method was unable to detect either intra-author or intra-author variations for the three authors in the study.

Because of the difference of the sample size, when trying to inspect the intra-author variation in Alice Sheldon's works the problem of overfitting occurs. Therefore, the intra-author variations will not be inspected employing supervised learning methods. The classification accuracy of the SVM analysis for discriminating between the three authors is 99.05%.

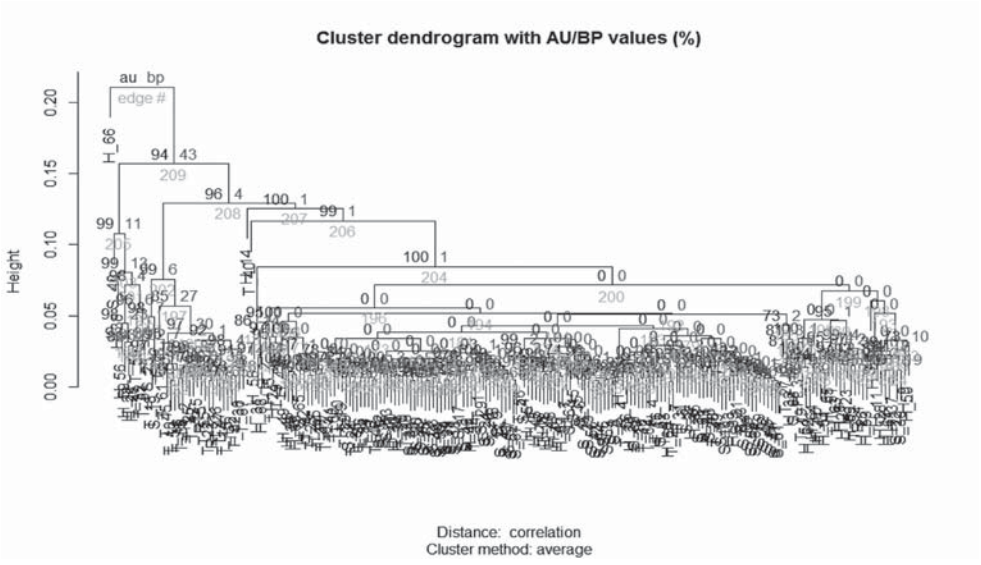


Figure 1: Cluster Dendrogram (Distance: Correlation, Cluster Method: Average) with Multi-Scale Bootstrap Resampling

Table 1: Results from Support Vector Machines (SVM) with a Cross Validation

	Alice Sheldon	Ernest Hemingway	Theodore Sturgeon
Alice Sheldon	**71**	1	0
Ernest Hemingway	0	**68**	1
Theodore Sturgeon	0	0	**70**

The second of the two supervised learning methods used random forests. Its results had a classification accuracy of 94.31%. From these results, we conclude that the supervised learning methods can successfully discriminate between the styles of Alice Sheldon, Hemingway, and Sturgeon.

Table 2: Results from Random Forests

	Alice Sheldon	Ernest Hemingway	Theodore Sturgeon	Class Error
Alice Sheldon	**67**	5	0	0.06944444
Ernest Hemingway	5	**62**	2	0.10144928
Theodore Sturgeon	0	0	**70**	0.00000000

By employing random forests, we can obtain proximity of data. Figure 2 gives a MDS plot of the random forests results based on proximity. The circles represent works by Alice Sheldon, the triangles represent those by Hemingway, and the crosses represent those by Theodore Sturgeon. Works that were misclassified are represented in their IDs. The plot shows evidence of three definite clusters for Alice Sheldon, Hemingway, and Sturgeon. Thus, by employing an MDS plot, inter-author variations can be detected. However, according to this plot, the cluster for Sturgeon is a little farther from the other two clusters. In this sense, Alice Sheldon's style is relatively similar to that of Ernest Hemingway, just as many literary critics have said. Moreover, based on Gini index, variables that are effective for discrimination are identified. For example, the possessive ending, singular nouns, and plural nouns are effective for this kind of discrimination.

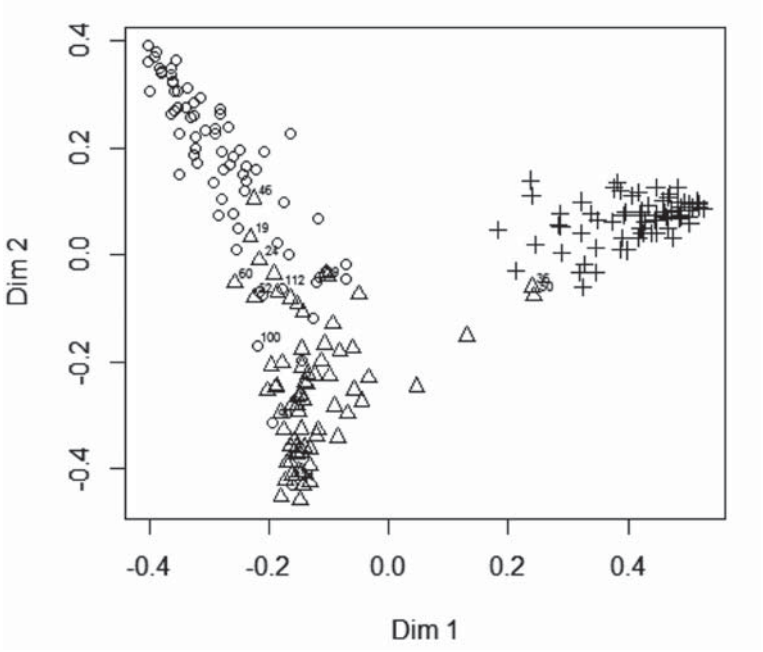


Figure 2: MDS Plot

Conclusion

This analysis makes two points. The first point addresses Le Guin's suggestion that the works by James Tiptree, Jr. and those by Raccoona Sheldon are different in terms of writing style. In fact, Alice Sheldon tried to change the content of her writing under her female pen name. However, as noted earlier, the statistical analyses employed for this study could not detect intra-author variations in Alice Sheldon's texts. In other words, Alice Sheldon failed to change her writing style and did not produce two different bodies of work to match her two differently gendered pseudonyms. The second point of this analysis considers the two kinds of supervised statistical analyses which did detect inter-author variations between the three authors' writing styles. Many literary critics suggest that James Tiptree's manner of writing is similar to Hemmingway's, in that both have a masculine writing style. However, our analysis did not prove that masculinity is the stylistic trait shared by Hemingway and Tiptree, though it does prove that their styles were similar. However, according to MDS plot, the cluster for Sturgeon is a little farther from the other two clusters. In this sense, Sheldon's style is relatively similar to that of Hemingway, just as many literary critics have said.

References

[1] Hirst, G., & Feiguina, O. G. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 22(4), 405-417.
[2] Hou, R., & Jiang, M. (2014). Analysis on Chinese quantitative stylistic features based on text mining. *Digital Scholarship in the Humanities*, 31(2), 357-367.
[3] Kimura, M. (2017). Quantitative Authorship Attribution of Two Authors and Evaluation of Variables — Alice Bradley Sheldon and Ernest Hemingway—, IPSJ SIG Technical Report, 2017-CH-113, 1-6.
[4] Larbalestier, J. (2002). *The battle of the sexes in science fiction*. Connecticut; Wesleyan University Press.
[5] Lefanu, S. (1988). *Feminism and Science Fiction*. Bloomington and Indianapolis; Indiana University Press.
[6] Silverberg, R. (1975). Who Is Tiptree, What Is He?, *Warm Worlds and Otherwise*. iv - x viii.

Annotation of 'Word List by Semantic Principles' Labels from the 'Corpus of Historical Japanese' Heian Period Series

– Trial Annotation on Tosa Nikki and Taketori Monogatari –

Masayuki Asahara (National Institute for Japanese Language and Linguistics), Nao Ikegami (Saitama University), Yutaka Hara (none), Sachi Kato (National Institute for Japanese Language and Linguistics), Tai Suzuki (none)

'Word List by Semantic Principles' (Bunrui Goi Hyo; hereafter WLSP) is a Japanese thesaurus which includes around 100,000 lexical items. The thesaurus offers two kinds of information: one regarding the syntactic feature (Rui), the other regarding the three-layered semantic features (Bumon, Chu-Koumoku, and Bunrui-Komoku). `Nihon Koten Taishou Bunrui Goi Hyo' (hereafter, historical-WLSP) was published in 2014, which is an enhanced version of WLSP, to cover historical Japanese. The semantic feature labels are based on the contemporary WLSP and extended to the senses that appeared only in historical Japanese. The book also includes a word list with the word frequency of 17 works from `Manyo-shu' in the Nara period to `Tsuretsure-gusa' in the Kamakura period. However, the word senses in the book are not disambiguated by the contextual information.
In order to resolve the word sense ambiguity issues, we performed a trial annotation of word sense disambiguation. We selected two samples -- `Tosa Nikki (Tosa Diary)' and `Taketori Monogatari (The Tale of the Bamboo Cutter)' -- from the `Corpus of Historical Japanese' of the Heian Period Series.

The annotation comprised the following three steps: (a) extract the contemporary WLSP labels, (b) extract the historical WLSP labels, and (c) choose the appropriate labels or (d) newly introduce WLSP labels.
In step (a), all possible contemporary WLSP labels are extracted for each morpheme. In step (b), the extended word senses are appended for each morpheme. In step (c), a human annotator chooses one sense from the all the possible word senses, and in step (d), introduces a new word sense when none of the word sense candidates is appropriate in the context.

We report the statistics of `Taketori Monogatari' and `Tosa Nikki', which have 12,757 and 8,208 morphemes, respectively. Annotations of WLSP labels cover 41.9% (5,348/12,757) and 42.4% (3,481/8,208) of the morpheme tokens.
From the 8829 (Taketori 5348, Tosa 3481) tokens, 1703 (Taketori 1060, Tosa 643) annotated tokens are from historical WLSP labels. Therefore, 19.3% of the annotated tokens are word senses only in historical Japanese.

Table 1: The statistics of the syntactic feature (Rui)

	1 Noun	2 Verb	3 Adj, Adv	4 Others	Total
Taketori	2318 43.3%	2252 42.1%	706 13.2%	72 1.3%	5348
Tosa	1710 49.1%	1272 36.5%	454 13.0%	45 1.3%	3481
Total	4028 45.6%	3524 39.9%	1160 13.1%	117 1.3%	8829

Table 1 shows the syntactic feature level of the two samples.
The MVR (Modifier-Verb Rate) of Taketori and Tosa are 31.3% and 35.6%, respectively.
To make a more meaningful comparison, we introduced separation by sentence types, such as 'descriptive' 地の文 , 'poem' 歌 , 'quote' 会話 , and 'letter' 手紙 , in Table 2. The descriptive part of the Noun Rate and MVR are nearly the same as that in the preceding research by Fujiike (2013) in which the rates are estimated by the POS information.

Table 2: The statistics of the syntactic feature (Rui) by sentence types

	1 体 Noun	2 用 Verb	3 相 Adj, Adv	4 他 Other	(subtotal)	Noun Rate	MVR
Taketori							
:descriptive	1147	1230	289	28	2694	42.5%	23.5%
:poem	60	49	13	2	124	48.3%	26.5%
:quote	1026	891	375	39	2331	44.0%	42.0%
:letter	85	82	29	3	199	42.7%	35.4%
(subtotal)	2318	2252	706	72	5348	43.3%	31.3%
Tosa							
:descriptive	1298	966	368	38	2670	48.6%	38.1%
:poem	267	202	48	2	519	51.4%	23.8%
:quote	145	104	38	5	292	49.8%	37.7%
(subtotal)	1710	1272	454	45	3481	49.1%	35.7%
Total	4028	3524	1160	117	8829		

Table 3 shows the statistics of the semantic features at the top level (Bumon). Interestingly, the semantic feature label distribution of the two samples are similar, although the syntactic feature label distribution of the two samples shows different types of writing styles.

Table 3: The statistics of the semantic features at the top (Bumon)

	1 関係 Relation	2 主体 Subject	3 活動 Action	4 生産物 Product	5 自然 Nature	9 枕詞 Poetic	Subtotal
Taketori	2335	577	1722	229	485		5348
	43.7%	10.8%	32.2%	4.3%	9.1%	0.0%	
Tosa	1540	360	1002	167	411	1	3481
	44.2%	10.3%	28.8%	4.8%	11.8%	0.0%	
Total	3875	937	2724	396	896	1	8829
	43.9%	10.6%	30.9%	4.5%	10.1%	0.0%	

Table 4 shows the top five frequent word senses in the most fine-grained semantic labels.
The distribution of the finest semantic features is slightly different between the two samples.

Table 4: The top five finest semantic features (Bunruikomoku)

Rank	Feature	The meaning of label	Taketori	Tosa	Total
Total	*	*	5348	3481	8829
1	1010	Relation-Thing-Pronoun	288 (5.3%)	163 (4.6%)	451 (5.1%)
2	1200	Relation-Existence-Existence	243 (4.5%)	136 (3.9%)	379 (4.2%)
3	3100	Action-Language-Linguistic Activity	271 (5.0%)	91 (2.6%)	362 (4.1%)
4	1527	Action-Operation-Reciprocating	150 (2.8%)	101 (2.9%)	251 (2.8%)
5	2000	Subject-Human-Human	127 (2.3%)	98 (2.8%)	225 (2.5%)

In our future work, we plan to annotate WLSP labels on other data in Heian periods and examine large-scale comparison among the samples.

Acknowledgements:

This work was supported by JSPS KAKENHI Grant Numbers 17H00917, 15K12888 and a project of Center for Corpus Development, NINJAL.

References:

[1] Yumi Fujiike, (2014) '品詞比率から見る中古和文テキストの特徴' (in Japanese),『日本語学会 2014 年度春季大会予稿集』
[2] Kokuritu-Kokugo-Kenkyusho, (2004) '分類語彙表', Bunrui Goihyo (Word List by Semantic Principles, Revised and Enlarged Edition) (in Japanese), Dainippontosho.
[3] Tatsuo Miyajima, Hisao Ishii, Seiya Abe, and Tai Suzuki, (2013) '日本古典対照分類語彙表' (in Japanese), Kasamashoin.

Building Networks and Creating Access in Early Modern Europe

Lisa Tagliaferri (City University of New York)

The literary contributions of marginalized writers have too often been neglected due to the privileging of an exclusive canon. As vernacular languages gained ground in an educated climate previously controlled by Latin literacy, more readers gained greater access to information and could become knowledge producers in their own right. One of the most prolific writers of 14th-century vernacular Italian was Catherine of Siena, a writer of nearly 400 extant letters, a book-length dialogue, and copied prayers. Though Catherine did work to learn to read Latin, she consciously and emphatically would choose Sienese Italian as her language of writing and speech. As a spiritual figure doing good work in the world, offering comfort, advice, and care to the prisoners, penitent women, and ill of Siena, she worked to include rather than exclude others. Leveraging network analysis and visualization, and considering how reader communities serve to provide access to information via the technology of writing, this paper seeks to show a historical example of the building of communities through literacy, authorship, and text.

This paper demonstrates the literary value of Catherine's writing through contextualizing her work among the corpus of contemporary texts, and analyzing her reception in the Renaissance period of Italy and England. This project situates Catherine within a larger textual community that adds to her authority as a writer with an influence beyond that of her language, culture, and century. Guided by humanistic inquiry concerning the interpretation of authorship and translation, and supported by quantitative digital methods, this work features an interdisciplinary approach to recover a marginalized writer. For this project, I have analyzed Catherine's body of texts against Dante's and Petrarch's major works, as well as the Middle English translation of her dialogue. This was done by my development of a script in Perl which cleaned up text files that I then analyzed through text mining in Python and Java-based topic modeling to see major topics across corpora. A substantial portion of the digital humanities analysis has been a complete network visualization of Catherine's correspondents across geographical Europe, and across class lines and religious affiliations. This work was done in R, JavaScript, and Python using D3 packages which export the information to visualizations that are ready for web interactive use. These interactive visualizations best express the reach of Catherine's textual production, and provide a comparison point to my traditional archival approaches that concern the history of medieval and early modern books. The website compendium to the dissertation, https://caterina.io, will continue to be developed to store relevant digital and interactive files, and to have a public-facing dimension to the research.

Proceedings of JADH conference, vol. 2017
Published by the Faculty of Culture and Information Science, Doshisha University
1-3 Tatara Miyakodani, Kyotanabe-shi, Kyoto 610-0394 Japan
Online edition: ISSN 2432-3144　Print edition: ISSN 2432-3187
Editor: Akihiro Kawase