JADH2016

Proceedings of the 6th Conference of
Japanese Association for Digital Humanities

“Digital Scholarship in History and the Humanities”

http://conf2016.jadh.org/
The University of Tokyo, September 12-14, 2016.

Hosted by:
JADH2016 Organizing Committee
under the auspices of the Japanese Association for Digital Humanities

Co-hosted by:
Historiographical Institute, The University of Tokyo(UTokyo)
Graduate School of Humanities and Sociology / Faculty of Letters, UTokyo
Center for Integrated Studies of Cultural and Research Resources, National Museum of Japanese History

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)

Co-sponsored by:
IPSJ SIG Computers and the Humanities
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: Analyzing Cultural Resources (Short papers)

Chair: Asanobu Kitamoto

Poster slam & poster session

Chair: Christian Wittern

Session 4: Textual Analysis (Long papers)

Chair: Toru Tomabechi

Session 5: Modeling and Digitization (Short papers)

Chair: Hajime Murai

JADH 2016 Organization

JADH 2016 Organizing Committee:

A. Charles Muller (University of Tokyo, Japan)
Makoto Goto (National Museum of Japanese History, Japan)
Shuhei Hatayama (University of Tokyo, Japan)
Akihiro Hayashi (University of Tokyo, Japan)
Yasufumi Horikawa (University of Tokyo, Japan)
Naoto Ikegai (University of Tokyo, Japan)
Hidetaka Ishida (University of Tokyo, Japan)
Tatsuo Kamogawa (University of Tokyo, Japan)
Masato Kobayashi (University of Tokyo, Japan)
Kiyonori Nagasaki (International Institute for Digital Humanities, Japan)
Hiroaki Nagashima (University of Tokyo, Japan)
Yusuke Nakamura (University of Tokyo, Japan)
Makoto Okamoto (University of Tokyo, Japan)
Masahiro Shimoda (University of Tokyo, Japan)
Akira Takagishi (University of Tokyo, Japan)
Noriyuki Takahashi (University of Tokyo, Japan)
Shogo Takegawa (University of Tokyo, Japan)
Toru Tomabechi (International Institute for Digital Humanities, Japan)
Kana Tomisawa (University of Tokyo, Japan)
Taizo Yamada (University of Tokyo, Japan), Chair
Shunya Yoshimi (University of Tokyo, Japan)

JADH 2016 Program Committee:

Hiroyuki Akama (Tokyo Institute of Technology, Japan)
Paul Arthur (Australian National University, Australia)
James Cummings (University of Oxford, 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)
Asanobu Kitamoto (National Institute of Informatics, Japan)
Maki Miyake (Osaka University, Japan)
A. Charles Muller (University of Tokyo, Japan)
Hajime Murai (Tokyo Institute of Technology, Japan)
Kiyonori Nagasaki (International Institute for Digital Humanities, Japan), Chair
John Nerbonne (University of Groningen, Netherlands)
Espen S. Ore (University of Oslo, Norway)
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)
Christian Wittern (Kyoto University, Japan)
Taizo Yamada (University of Tokyo, Japan)

Time Table

September 12, Day 0

13:00-14:30	Workshop •Management of Japanese Character Information and its Application •IIIF (International Image Interoperability Framework): Recent situation
15:00-18:00	Pre-conference symposium

September 13, Day 1

9:10-	Registration
9:45-10:00	Opening
10:00-11:30	Session 1: Texts and Database (Long papers)
11:30-11:50	Break
11:50-12:50	Session 2: History and Digital (Long papers)
12:50-14:20	Lunch Break
14:20-15:20	Keynote Lecture
15:20-15:30	Break
15:30-17:00	Session 3: Analyzing Cultural Resources (Short papers)
17:00-17:10	Break
17:10-18:40	Poster slam & poster session
19:00-	Reception

September 14, Day 2

9:30-11:00	Session 4: Textual Analysis (Long papers)
11:00-11:20	Break
11:20-12:50	Plenary panel session1
12:50-14:20	Lunch Break --- JADH AGM
14:20-15:40	Session 5: Modeling and Digitization (Short papers)
15:40-16:00	Break
16:00-17:30	Plenary panel session2
17:30-17:50	Closing

Pre-Conference Symposium

15:00-15:10 Opening

(Hiroshi Kurushima and Masahiro Shimoda)

15:10-16:10

16:10-16:20 Break

16:20-16:50

- Academic Assets and Digital Archives
  Noriko Kurushima (The University of Tokyo)

16:50-17:20

- Making Database of City Life from Genre Paintings - Persons' Database of 16th Century Rakuchu-rakugai-zu Byobu (Scenes In and Around Kyoto Screens)
  Michihiro Kojima (National Museum of Japanese History)

17:20-17:40 Break

17:40-18:00 Panel discussion including audience

Chaired by Makoto Goto (National Museum of Japanese History)

The Humanities, the Liberal Arts and the University in a Digital World

Peter K. Bol
Harvard University, USA

Abstract

What is the role of the humanities in education and why are the humanities central to the liberal arts? The job of the humanities is to remember our past, both its best and its worst, when it is easier to forget; to push us reflect on ourselves and question our present when it is easier to go along. Above all else, the humanities continue our predecessors' efforts to create and sustain civilization. They remind us that, as Confucius said, "Learning without thinking is to deceive oneself; thinking without learning is to endanger oneself 學而不思則罔, 思而不學則殆 ." When learning is treated as acquiring skills employers can use and thinking is reduced to following simplistic ideologies, the humanities offer the antidote.

The digital world gives the humanities new possibilities to help us learn, reflect, and create. Its tools allow us to see more, to think more clearly, and to communicate across cultures. We need to consider how the humanities can embrace these tools and skills without losing sight of its mission and without forgetting its past.

Biography

Peter K. Bol is the Vice Provost for Advances in Learning and the Charles H. Carswell Professor of East Asian Languages and Civilizations. As Vice Provost (named in 2013/09) he is responsible for HarvardX, the Harvard Initiative in Learning and Teaching, and research that connects online and residential learning. Together with William Kirby he teaches ChinaX (SW12x) course, one of the HarvardX courses. His research is centered on the history of China's cultural elites at the national and local levels from the 7th to the 17th century. He is the author of "This Culture of Ours": Intellectual Transitions in T'ang and Sung China, Neo-Confucianism in History, coauthor of Sung Dynasty Uses of the I-ching, co-editor of Ways with Words, and various journal articles in Chinese, Japanese, and English. He led Harvard's university-wide effort to establish support for geospatial analysis in teaching and research; in 2005 he was named the first director of the Center for Geographic Analysis. He also directs the China Historical Geographic Information Systems project, a collaboration between Harvard and Fudan University in Shanghai to create a GIS for 2000 years of Chinese history. In a collaboration between Harvard, Academia Sinica, and Peking University he directs the China Biographical Database project, an online relational database currently of 360,000 historical figures that is being expanded to include all biographical data in China's historical record over the last 2000 years.

[Keynote Lecture]

Credit where credit is due: how digital scholarship is changing history in the English-speaking world and what the American Historical Association is doing about it

Seth Denbo, Ph.D.
American Historical Association, USA
sdenbo@historians.org

"The context of historical scholarship is changing rapidly and profoundly." With these words the American Historical Association launched its intervention into the problem of evaluating digital scholarship by historians.

As historians, we are conducting our scholarship (research, teaching, writing, publishing) in a world that is changing rapidly. In every stage of historical research the digital context of our work is transforming what we do. Teaching is also being refigured by the use of digital tools and methods. These methodologies are no longer the preserve of a small minority of digitally-trained historians. Even scholars with limited technological skills use the web for finding primary and secondary sources, doing basic computational analyses, and even publishing online. The use of digital technologies gives us new ways to approach our traditional questions, provides more varied forms of expressing ideas, and allows us to reach new audiences.

While the conduct of our historical work has changed in many ways, we lag behind in evaluating scholarship using non-traditional methods. Disciplinary imperatives limit forms of acceptable publication to traditional outputs—journal articles and books. The peer review that underpins the entire process of scholarly publication often does not occur when work is published online. Lacking peer review mechanisms, many departments are reluctant to open their requirements for tenure and promotion to these new approaches and formats.

Developments in digital history are changing what we can express about the past. I will explore how through looking at some exemplary uses of these approaches in the English-speaking academic world. Digital history is not a new phenomenon. Economic and social historians realized the power of computational tools for analyzing large-scale data as long ago as the 1970s. This work suffered from making promises that were impossible to deliver on, and was overtaken by a cultural and linguistic turn in the wider discipline of history. But it provided a foundation for conceptualizing how large-scale data sets covering broad swathes of historical time could become an important methodological approach for our discipline.

Today the work of digital historians is much more varied and immersed in existing paradigms. It is more of a set of lenses for viewing sources than a separate field within our discipline, but those lenses are highly varied. Some provide very close and detailed interpretations of a small number of sources, while others look at vast amounts of data to paint a picture of change over time. Other approaches are primarily about historical education, both in and out of the classroom. In looking at these projects my paper will examine how they contribute to the scholarly conversation in their field, explore some of the challenges they present to traditional modes of scholarship, and discuss issues related to evaluating them for professional credit.

[Plenary panel session 1]

Three Databases on Japanese History and Culture: an Editing Experience

Charlotte von Verschuer
École Pratique des Hautes Études, France

Abstract

I will present three internet databases related to Japanese history and culture that I have co-edited and co-authored.

Online Glossary of Japanese Historical Terms 日本史グロッサリー・データベース　or:
On-line Glossary of Japanese Historical Terms　応答型翻訳支援システム　.

The Online Glossary of Premodern Japanese Historical Terms is one of the sub-projects of the Japan Memory Project (JMP), designed and created with the support of the Ministry of Education, Culture, Sports, Science and Technology (COE, 2000 – 2004), the Japan Society for the Promotion of Science (Grant-in-Aid for Scientific Research, 2005-2008) and a number of foreign scholars.
Project Director: Ishigami Eichi, Director of the Japan Memory Project (2000-2008);
Members of the Advisory Committee:
Martin Collcutt (Princeton University), Kate Wildman Nakai (Sophia University), Joan Piggott (University of Southern California), Detlev Taranczewski (Universität Bonn), Ronald P. Toby（University of Illinois, Urbana Champagne), Hitomi Tonomura (University of Michigan), Charlotte von Verschuer（École Pratiques des Hautes Études), Willy Vande Walle（Katholieke Universiteit Leuven. All other Project members, Editorial staff, and Editorial assistants are listed on the site.
The purpose of this glossary is to select and list major existing translations for Japanese historical terms and to make them available over the internet as a tool for assisting in the translation of Japanese primary sources. The glossary consists of more than 25,000 entries. Instead of giving set translations or any English standard terms, the glossary, as a special feature, provides a variety of translations for the same technical term and gives, for each translation, the author name and publication. The glossary is drawing these translations from over 70 works written in English, French, and German.

Dictionary of Sources of Classical Japan / Dictionnaire des sources du Japon classique
欧文日本古代史料解題データベース　(Online Draft Version" December 2004)

Book Version:	Dictionnaire des sources du Japon classique/Dictionary of Sources of Classical Japan, Paris: College de France, 2006; distribution: De Boccard: http://www.deboccard.com/
Editors:	Joan Piggott, University of Southern California Ineke Van Put, Catholic University of Leuven Ivo Smits, Leiden University Charlotte von Verschuer, École Pratiques des Hautes Études Michel Vieillard-Baron, Institut National des Langues et Civilisations Orientales (INALCO)
Co-editors:	Ishigami Eiichi, Historiographical Institute, The University of Tokyo (Shiryô Hensanjo) Yoshida Sanae, Historiographical Institute, The University of Tokyo (Shiryô Hensanjo) Horikawa Takashi, National Institute of Japanese Literature (NIJL; Kokubungaku Kenkyû Shiryôkan) / Tsurumi University
Advisors:	Araki Toshio, Senshû University Sano Midori, Gakushuin University Brian Ruppert, University of Illinois, Urbana-Champaign

Tabuchi Kumiko, National Institute of Japanese Literature (NIJL; Kokubungaku Kenkyû Shiryôkan)
Kikuchi Hiroki, Historiographical Institute, The University of Tokyo (Shiryô Hensanjo)

Collaboration: National Institute of Japanese Literature (NIJL; Kokubungaku Kenkyû Shiryôkan); Centre de recherches sur les Civilisations chinoise, japonaise et tibetaine (UMR-CNRS, EPHE, College de France, Universite de Paris 7)

Support:
- Japan Memory Project (JMP) at the Historiographical Institute, The University of Tokyo (Shiryô Hensanjo)
- École Pratique des Hautes Études (EPHE), Section des Sciences Historiques et Philologiques,

Traditional Agricultural Techniques: A Glossary in French-English-Chinese-Japanese (Grains and Horticulture) Preliminary Version 2013
農業技術用語集：仏・英・中・日（穀類）２０１３年暫定版（インターネット・データベース）
http://labour.crcao.fr
New Title (November 2016):
Dictionary of Traditional Agriculture: English-French-Chinese-Japanese
Dictionnaire de l'agriculture traditionnelle: français-anglais-chinois-japonais
法英汉日传统农业辞典
伝統農業技術：英日中仏用語辞典
Editors (2016): Cozette Griffin-Kremer (Conservatoire National des Arts et Métiers CNAM), Guoqiang Li (Paris West University), Perrine Mane (Centre National de Recherches Scienfiques CNRS), Charlotte von Verschuer (EPHE)
Authors:
Yoshio Abe (École des Hautes Études en Sciences Sociales EHESS), Carolina Carpinschi, Cozette Griffin-Kremer, Guoqiang Li, Perrine Mane, Francois Sigaut (EHESS, CNAM), Eric Trombert (CNRS), Charlotte von Verschuer
Advisors: Michiaki Kono (Kanagawa University, Yokohama), Takeshi Watabe (Tokai University, Tokyo), Yin Shaoting (Yunnan University, China)
Webmaster: Philippe Pons (EPHE)
Technical Management: Elise Lemardelée (EPHE), Yves Cadot (Université de Toulouse)
Publisher: East Asian Civilisations Research Centre (CRCAO: EPHE, CNRS, Université Paris- Diderot, College de France)
Date of publication: 2009, 2013, 2016
Collaboration: Research Group on the Comparative History of Agricultural Technology
Support: Fondation pour l'étude de la langue et de la civilisation japonaises (Fondation de France), Paris; Fukushima Prefectural Museum, Japan; China Agricultural Museum, Beijing; Institute of Botany (Chinese Academy of Sciences), Beijing, China.

- In contrast to a dictionary, this glossary is not meant to be exhaustive. It provides a selection of technical terms, deliberately excluding most generic terms. The glossary emphasizes technical specifics. We hope that it will enable users to avoid some common errors of translation by refining the meanings given for equivalent items.
- With the exception of words noted as older (ANC.), the terms listed are contemporary.
- The glossary covers traditional agricultural techniques, as they were practiced around the world up to this day. Terms that arose after industrialization have been excluded. (For these terms, the user can refer to industrial machine and product catalogues.)
- This glossary can contribute to safeguarding a wealth of technical information and knowledge about biodiversity, potentials for food production and wise utilization of resources and energy.
- The glossary highlights cultural differences: many technical terms have no equivalent in another cultural area. (The symbol @ attached to a word means that the term is specific to a particular language.)

- The entries are arranged by thematic category, so a search can be carried out either by word or by thematic category. Each entry has a window in which users can enter their own comments.

The Contents:

The *Dictionary* contains technical terms of agricultural traditions in a thematic arrangement. Many terms are documented by pictures. The Draft Version published in 2013 comprises the techniques of grain cultivation, vegetable and fruit agriculture, providing terms for agricultural operations and tools. The *Dictionary* is arranged in eleven thematic categories with a total of about ten thousand entries, covering: Tillage, Water Management, Soil Improvement, Sowing, Harvesting, Threshing-Degraining, Cereal Grains, Fruit and Vegetables, Plant Morphology, Fields and Systems, and Horticulture. The parallel presentation of English, French, Chinese and Japanese terms will shed light on the technical and cultural differences between the various linguistic areas. The *Dictionary* comprises the basic techniques, both traditional and contemporary. It does however not include the variants that involve the use of fuel, of chemicals, and of biotechnology, as these terms can be found on commercial catalogues. The project espouses the need to protect natural resources and preserve rural cultural heritage.

Perspective:

In an age of concern over saving the environment and bio-diversity, it seems timely to provide information about agricultural techniques that support this aim. In light of the high stakes involved in climate change, economic globalization and the industrialization of agriculture, traditional agricultural techniques deserve to be considered as a universal asset of humankind. The *Dictionary* has first been launched on-line in 2009. It is continuously expanding and will cover fields other than grains, vegetable and fruit agriculture, such as cattle husbandry, viticulture, sylviculture etc.

Aim: With the world wide concern for global Food Security, research on agricultural techniques is progressing in European countries as well as in China and Japan. It is time to provide a working tool for translations and international communication. It goes without saying that the general language dictionaries do not provide precise enough information in the field. The *Dictionary* should be used for translating technical works and catalogues. The *Dictionary* will enhance the study of environmental ecology and be the safeguard of rural heritage. It should promote research and fieldwork by graduate students and curators and, last but not least, it encourages a dialogue among the specialists of various countries.

Biography

Charlotte von Verschuer is Professor of Japanese history at École Pratique des Hautes Études in Paris. Born in Bonn, Germany, she did her school education in Brussels at the European School, in Belgium. She then studied Japanese at the International Christian University in Tokyo, Japanese and Chinese languages, as well as Asian art history at Bonn University in Germany, and graduated in Japanese studies at the Institut National de Langues Orientales (INALCO University) in Paris. Thereafter she spent two years as a Japanese Government scholarship fellow at the Institute of History (*Kokushi kenkyushitsu*) at The Tokyo University under the guidance of Tsuchida Naoshige with his Ishii Masatoshi, and also spent eight months as a trainee at the Taiwan Palace Museum in Taibei, and continued her Ph.D. studies in Paris, Ecole Pratique des Hautes Etudes (EPHE) under the guidance of Francine Herail. She received her Ph.D. in Oriental Studies at INALCO University with her thesis on ‘8th-9th Century Official Relations between Japan and China’, and an other Ph.D. in History at Paris EPHE with her thesis on ‘The Economy of Ancient Japan’. She was associate researcher at Centre National de Recherches Scientifiques (CNRS) before becoming Professor of Ancient and Medieval History of Japan at EPHE in 1995, at the East Asian Civilisations Research Centre (CRCAO). Her publications in French, German, English, and Japanese include: - *Across the Perilous Sea: Japanese Trade with China and Korea from the seventh to sixteenth Centuries*, translated from French by Kristen Lee Hunter, Ithaca (New York), Cornell University Press, 2006; and - *Rice, Agriculture, and the Food Supply in Premodern Japan*, translated and edited by Wendy Cobcroft, London, Needham Research Institute Monograph Series, London, New York, Routledge, 2016.

[Plenary panel session 1]

Intellectual Networks in Tokugawa Japan: the beginnings of the Edo Japan Database

Bettina Gramlich-Oka, Ph.D.
Sophia University

Abstract

The project is a historical network analysis of the Tokugawa period (1600–1868). Our principal actor is the scholar Rai Shunsui (1746–1816) and his many records. Shunsui's diary, spanning over thirty-five years, his correspondence, and many other records are rich in information regarding the wide intellectual network that Shunsui nurtured and that extended all over Japan.
The project offers thus a novel approach in that it is not simply an intellectual biography but grounded in the notion that intellectual interactions among scholars of the Tokugawa period are much better described by the analogy of a network. Their correspondence, meetings and sharing of objects and manuscripts will help us to understand better the actual working of the various levels of state administration, in which the scholars were involved. Therefore, intra-territorial and inter-territorial networks are keys to understanding how political reforms were discussed and implemented in Tokugawa Japan. In more concrete terms, this project will investigate the network of Rai Shunsui in order to document the intellectual environment of the late Tokugawa reforms in time and space by setting up a geo-database (GIS) containing the data collected from a broad variety of sources.

Biography

Bettina GRAMLICH-OKA holds a Ph.D. (University of Tübingen, Germany) in Japanese history. She is a professor for Japanese history at Sophia University, Tokyo, where she teaches courses in women history, Edo society, and upper level courses implementing "reacting to the past" pedagogy. Her main publications are *Thinking Like a Man: Tadano Makuzu* (Brill, 2006; in Japanese 2013), *Economic Thought in Early Modern Japan* (Brill, 2010; in Japanese 2013), and is currently working on intellectual networks, marriage and adoption practices in the Edo period. In 2014 she became the editor and since 2016 the Chief Editor of *Monumenta Nipponica*. Since 2010, she is the leader of the research unit "Network Studies" in the Institute of Comparative Culture of Sophia University (network-studies.org). Part of the project is the development of the relational database introduced here.

The Kanseki Repository: A new online resource for Chinese textual studies

Christian Wittern (Kyoto University)

Introduction

The Kanseki Repository (KR) has been developed by a research group at the Institute for Research in Humanities, Kyoto University under the leadership of Author(s). It features a large compilation of premodern Chinese texts collected and curated using firm philological principles based on more than 20 years of experience with digital texts. Among its unique features is the fact that the texts can be accessed, edited, annotated and shared not only through a website, but also through a specialized text editor, which thus morphes into a powerful workspace for reading, research and translation of Chinese texts. The Kanseki Repository includes all texts in the Daozang and Daozang jiyao and a large collection of Buddhist material, including all texts created by the CBETA team, where applicable enhanced through the inclusion of recensions from the Tripitaka Koreana, in addition to a large selection from general collections like *Sibu congkan* and *Siku quanshu*.
The source texts of the Kanseki Repository are available at @kanripo on the website github.com. These texts are displayed at www.kanripo.org and also used in the Emacs Mandoku (see www.mandoku.org) package.
This presentation will outline the main considerations for creating this repository of texts and its associated tools and methods. This includes

- *Philological foundations*
- *Basic technologies*
- *Cooperative and collaborative research*

These points are further discussed below

Philological foundations

In a seminal article, the Swiss scholar Hans Zeller[1] emphasised the fact that all scholarly editing should make a clear distinction between the **record** of what is transmitted and the scholarly **interpretation** thereof. While this distinction is blurry at times, it has informed the design of the *Kanseki Repository*, which arranges the editions of a text it represents into those that strive to faithfully reproduce a text according to some textual witness ('record') and those that critically consider the content and make alteration to the text by adding punctuation, normalizing characters, collating from other evidence etc. ('interpretation').

Basic technologies

Git and GitHub

The distributed version control software git is used as a low-level transportation layer and maintenance technology. It allows users to download texts and upload revised versions, create their own versions and keep track of revisions. Github is a commercial web services based on git, that adds social-networking functions and cloud-services.

Emacs

Emacs is the main user interface for users that require a sophisticated and advanced editing environment. On top of the Emacs package "Org mode" has an extension been developed that adds additional functionality that facilitates interaction with the digital archive.

Web interface at www.kanripo.org

This website provides access to the texts, including full-text search, display of transcribed text and facsimile(s) of different editions. Users can log in using their Github credentials and get access to more advanced functions such as selecting lists of text of special interest, advanced sorting functions by text category or date as well as cloning of texts to the Github user account and editing on site. The site went into testing mode in October 2015 and is scheduled to a first public release in March 2016.

Towards a platform for text-based Chinese studies

All modes of interaction described above are based on the distributed version control system `git`, using the Github site as a 'cloud storage'. However, in addition to providing storage, Github also provides a feedback mechanism through "pull-requests", where users can flag corrections to a text for the `@kanripo` editors to consider for inclusion in the canonical version, thus making it available to all users.

The model outlined here is extensible and allows other developers of websites related to Chinese studies to access the same texts, and provide specialized services to the user, for example by enhancing the text through NLP processing. These enhanced versions can be saved ("committed" in `git` language) in the same way to the users account and are then also visible to the client programs described here[1].

This will open the door to an open platform of texts for Chinese studies, where the texts of interest to the users form the center of a digital archive, with different services and analytical tools interacting and enhancing it. The user, who makes a considerable investment in time and effort when close reading, researching, translating and annotating the text, never loses control of the text and does not need to worry about losing access to it when one of the websites goes offline.

By providing versioned access to the texts in question, it is also possible to make any analytical results reported in research publications reproducible[2] by indicating the additional tools and processes needed, ideally also in a Github repository in the same ecosystem.

The aim is not just to provide a static, completed, definitive edition of a text, but as fertile a ground as possible for the interaction between the text and its readers, hopefully improving both through this process.

References

[1] Hans Zeller, "Befund und Deutung - Interpretation und Dokumentation als Ziel und Methode der Edition", in: G. Martens and H. Zeller (ed.), *Texte und Varianten : Probleme ihrer Edition und Interpretation*. München, 1971, p. 45-89, translated as "Record and Interpretation: Analysis and Documentation as Goal and Method of Editing" in: Hans H. W. Gabler, G. Bornstein, and G. B. Pierce (ed.), *Contemporary German Editorial Theory*, Ann Arbor 1995, p. 17-58.

[2] Vikas Rawal, "Reproducible Research Papers using Org-mode and R: A Guide", at https://github.com/vikasrawal/orgpaper [accessed 2016-05-04]

[1] A “shadow” of the texts in the @kanripo account in a format suitable for text mining have been made available for specialized processing in @kr-shadow (http://github.com/kr-shadow). These texts will be updated from the master-branch of a corresponding text in @kanripo.

Migration, Mobility and Connection: Towards a Sustainable Model for the Preservation of Immigrant Cultural Heritage

Paul Arthur, Jason Ensor (Western Sydney University), Marijke van Faassen, Rik Hoekstra, Marjolein 't Hart (Huygens ING), Nonja Peters (Curtin University)

All over the world migrants have influenced and changed the cultures of the countries where they have settled, and they have built new communities that have retained connections, to differing degrees and by various means, with their original homelands. The multiple traces that they have left in official and unofficial documents potentially provide a rich resource for supporting and celebrating a sense of identity within such communities and for capturing and maintaining their histories. The gathering and preservation of these histories are also fundamentally important for enabling research on immigrant cultural heritage and thereby contributing to deeper understanding of cross-cultural and multicultural issues in an era of unprecedented global movement of people away from their homelands. In the case of migrants, collecting information that can provide relevant data is complicated by the fact that at least two countries are involved, with different laws, policies and conventions for data storage and access, and also in most cases, different languages.

In this project between two countries, via close collaboration the Digital Humanities Research Group at the University of Western Sydney and the Huygens Institute for the History of the Netherlands in The Hague, sets up processes for overcoming barriers such as these that have stood in the way of cross-national research on migrant lives in the past.

The importance of cultural heritage to national economies and social capital is widely recognised. In 2014 the Council of the European Union adopted the 'Conclusions on Cultural Heritage' confirming cultural heritage as 'a strategic resource for a sustainable Europe'. The 'Conclusions' recognised the role of participatory governance in 'triggering new opportunities brought by globalisation, digitisation and new technologies which are changing the way cultural heritage is created, accessed and used'. [1] It is these new opportunities that this 'Migration, Mobility and Connection' project responds to.

Documents and evidence of the history of migration are spread very widely, and in most cases have been almost entirely inaccessible for research purposes in the past in Australia. This project is a study on Dutch-Australian mutual cultural heritage. Its aim is to begin the process of finding, assembling and organising into accessible and searchable formats, information in selected key archival records, in both Australia and The Netherlands, relating to Dutch emigration to Australia. The project is conceptualised as a pilot that addresses difficulties faced by transnational collaboration of this kind and proposes ways of overcoming them. It will work through archival and custodial challenges in the discovery, collection, preservation and content management of traces from the past and propose new digital approaches that may lead to solutions. While the initial focus will be on migration, in the context of the maritime and mercantile history that the Netherlands shares with Australia, the project aims to establish a model that can be utilised for further Netherlands–Australian mutual heritage work and, potentially, for other immigrant groups.

Joint activity is underway to design a database for the project that integrates data in Australia with data in The Netherlands. Three digitised datasets contain representations of migrant travels: (a) Netherlands database (registration cards); (b) National Archives of Australia database (casefiles from several series); and (c) Nominal rolls / ships' passenger lists (representing a high percentage of digitisation in the National Archives of Australia). Items (a) and (b) are to be used for the data backbone; item (c) can be used for a more geographic visualisation (migrant mobility between the Netherlands and Australia and vice versa) and enrichment of the data backbone. The three datasets are different sources of information about the same people and voyages; they can therefore be used to determine where each of them has structural gaps (if any) and make it

[1] See http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52014XG1223(01) (accessed 25 May 2016)

possible to produce a more detailed estimate of the numbers of people that migrated and the way they travelled.

Reorganising a Japanese calligraphy dictionary into a grapheme database and beyond: The case of the *Wakan Meien* grapheme database

Kazuhiro Okada
ILCAA, Tokyo University of Foreign Studies

Hiragana, a Japanese moraic script, had long had a variety of letters before standardisation in 1900. Our knowledge of the history of hiragana has been deepened from the historical relationships to distinguishing letter usages of letters. However, little of our knowledge has been translated into machine-readable form. Consequently, the 1000-year-long tradition of hiragana before 1900, or older hiragana, is still left underrepresented in the computational world. This paper will address issues concerning reorganising a Japanese calligraphy dictionary, *Wakan Meien*, into a grapheme database, and discuss its further use as a knowledge database of the older Japanese writing system.

Wakan Meien is a calligraphy dictionary specialising in hiragana materials, compiled by To Koei (birth and death dates unknown) and published in 1768 (Fig. 1). Hiragana developed from cursivised Chinese characters. Today, it consists of 48 letters, whereas before the Meiji period, it had many more. Its cursivised origin makes it difficult to distinguish between levels of cursivisation, although some are distinguishable. *Wakan Meien* is one of the earliest kana dictionaries, and was compiled to meet growing demand by calligraphy students. The dictionary is unique, in that it presents examples grouped by similarity of shapes and not by genetic relationship. Genetic classification is a method that groups according to the source of cursivisation, and is still commonly used in later dictionaries (Fig. 2).

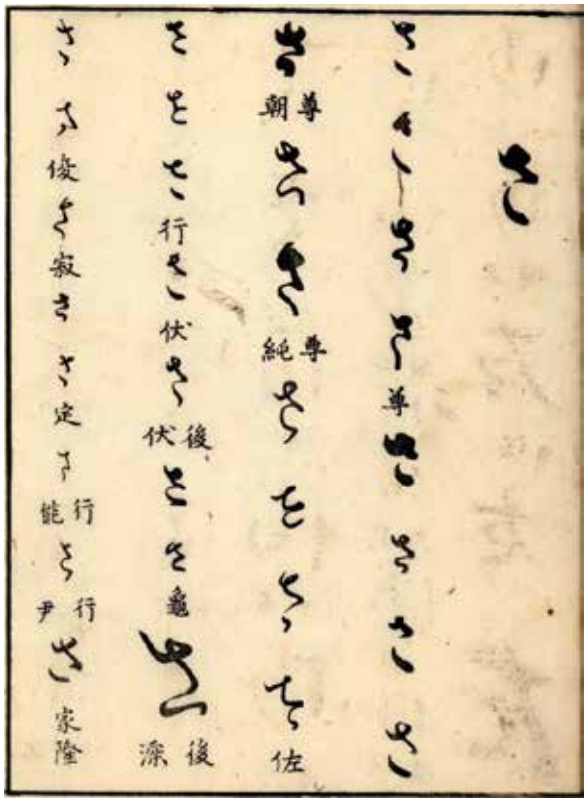

Figure 1. *Wakan Meien*

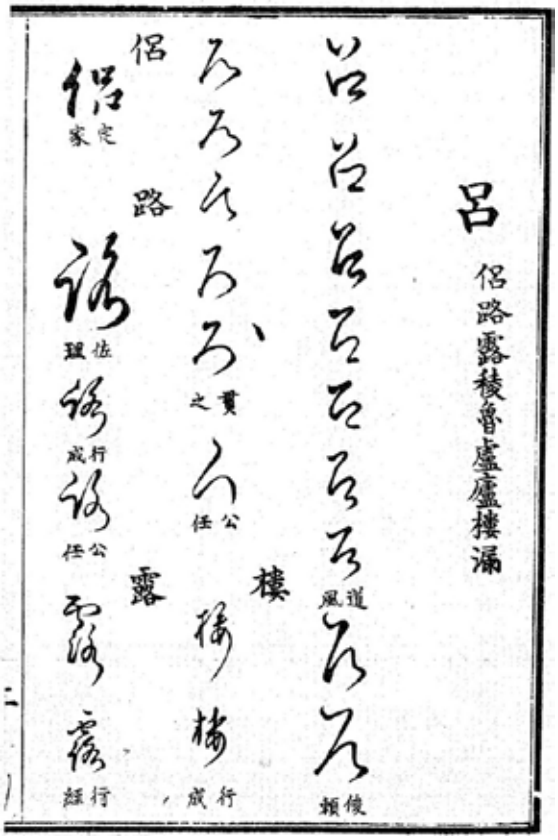

Figure 2. An example of genetic classification in *Kana Ruisan* (Sekine Tametomi, 1768. Holding of NDL Digital Collection)

Figure 3. Views of Sections, Groups, and Examples

The organisation of *Wakan Meien* surpasses later dictionaries with regards to grapheme representation, the basic units of a writing system. Genetic classification is generally well regarded amongst academics as an objective method, based on the fact that relationships between hiragana and cursivised Chinese characters are philologically clear, and that, further, it does not refer to the researcher's distinction between graphemes. However, deep understanding of the distinction between graphemes is essential, in order to ensure consistent computational encoding, such as Unicode. Conversely, the organisation of *Wakan Meien* means that these groups of examples correspond to distinction of graphemes (Okada, 2016). Building a grapheme database from genetic classified dictionaries involves complex and uncertain differentiation: Thus, it is necessary to build a grapheme database from attested graphemes, including those of *Wakan Meien*, for example.

The dictionary appears not to be well organised. It collates examples by order of Iroha, a common Japanese mnemonic of hiragana. Then, examples are ordered by similarity of shapes: a group of more Japanised shapes includes solely similar shapes, whilst that of less Japanised — in other words, retaining more of the Sinitic original — shapes includes many more variations in cursivisation, from barely to largely. These groups are not strictly ordered, other than that more Japanised shapes tend to appear first. The source Chinese character is not considered in the ordering. This ordering may give an impression for readers used to genetic classification, that examples are not well organised.

Reflecting that structure, the database recognises the following 3 entities: Sections, Groups (of examples), and Examples (Fig. 3). In the database, each group carries the possible variations, e.g., the distinction between graphemes. Considering that the original work is not strictly structured, the database presents relationship between entities rather than structure of them like tree. In addition, those entities have their own properties, such as heading images in Sections, source characters in Groups, and locations and authors of the examples in Examples. Some of these properties, source characters and authors of the examples to name a few, may have two or more sub-properties.

As will be discussed later, the database will be offered as a reference of older hiragana. This nature requires that points of reference to groups should not be excessively altered. Substantial updates to Groups thus should not impact existing references to them, but be made through creating new ones. This means that Examples can have relationship between two or more Groups.

A document(-oriented) database is employed to manage such data. Major advantages in employing document databases, compared to relational databases, include that it allows structured data to be stored as they are. Whilst relational databases can also manage such data after normalisation, recalling the loose structure of the original work, allowing it at scheme level would help development of better schemata.

The database will be provided as a reference source of older hiragana. It will include an educational purpose, in learning to read materials that are written in older hiragana, manuscripts of Japanese Classic for example, as well as a resource in corpus building, either in the form of linked data, or simply a link in an HTML page. First, with recent advances in mobile applications for learning older hiragana, such as 'the Hentaigana app' by the UCLA-Waseda alliance and 'KuLA' (Kuzushiji Learning Application) by Osaka University, it is expected to increase the broader popularity of older hiragana. The database will provide supplemental materials for learners. Second, it will be a reference for corpus building. Whilst older hiragana will be registered for Unicode in the near future, its current specification declares that it will not deal with the detail of distinctions between graphemes. Hence, building corpora in such a way that allows such a distinction must rely on other resources. The database will provide a reference for detail via either graphemes or actual examples, using stable IRI (Internationalized Resource Identifier). Moreover, accumulation of those links to the database, or links to other databases, will enable the formation of a knowledge database of older hiragana, and further the entire Japanese writing system, comprehending its structure and history with firm examples.

Reference

[1] Okada, Kazuhiro. 2016. *Wakan Meien* ni okeru hiragana jitai ninshiki [Hiragana grapheme awareness in *Wakan Meien*]. Paper presented at the 2016 Spring meeting, the Society for Japanese Linguistics, Gakushuin University, Tokyo, May 2016.

Enhancing ISO Standards of temporal attributes in information systems for historical or archaeological objects

Yoshiaki Murao (Nara University), Yoichi Seino, Susumu Morimoto (Nara National Research Institute for Cultural Properties), Yu Fujimoto (Nara University)

In this paper we attempt to implement the temporal attributes of historical or archaeological objects in information systems by enhancing ISO 19108 standards.

There is no doubt about the importance of temporal attributes for humanities. And the standardization of temporal attributes is also important to utilize IT for integrating or exchanging data of humanities. From standardization's point of view for the digital expression of temporal attributes, there are some discussion points about the characteristics of them, which is from semantic concepts to encoded formats. CIDOC/CRM (its official standard is ISO 21127) defines the semantic model of heterogeneous cultural heritage information, and contains the temporal element as "E2 Temporal Entity", "E51 Time Span", "E62 Time Primitive" and so on. These semantic common class definitions are valuable for the application area of cultural heritage and museum documentation, however CIDOC/CRM does not approach to build a general concept of temporal attributes for information resources of humanities, nor cover the encoding specifications of each class. A standardized implementation specification for E15 or E62 of CIDOC/CRM has to be required, and our study is positioned there.

There are currently two major international standards for common temporal attributes. One is ISO 8601 titled "Data elements and interchange formats – Information interchange – Representation of Dates and Times", which is based on Gregorian calendar and Coordinate Universal Time (UTC). For example, the format of "2011-03-11" is conformed to ISO 8601. It is widely used in representing date or time on information system. Although it provides convenient representation forms for recent events or activities, it is not suit for describing historical events, as they sometimes cannot be applied Gregorian calendar, and are required to use complex temporal expressions.

The other is ISO 19108 "Geographic information – Temporal schema", which defines the schema in order to implement many types of calendars or eras. It also defines the ordinal era to support the Jurassic period or the Cretaceous period, that are classified the order of periods. In contrary to ISO 8601, it can potentially support complex temporal expressions. In addition, ISO 19108 links to other encoding specifications in the same standards family. ISO 19118 "Geographic information – Encoding" provides basic encoding rules and ISO 19136 "Geographic information – Geography Markup Language" provides a practical encoding specification based on XML.

"<gi:date8601>2011-03-11</gi:date8601>" and "<gi:ordinalPosition idref="NaraPeriod"/>" are core parts of XML encoded examples conformed to ISO 19108 and ISO 19118.

ISO 19108 defines the common data model for temporal characteristics with varieties of temporal expressions. However, it is not sufficient to express the temporal attributes of humanities' objects, especially for historical or archaeological objects. Because these objects sometimes cannot be assigned the year of existence in any calendar, they sometimes use originally defined period or era, whose start or end time of their time span sometimes cannot specify clearly.

Then, we considered following cases of expressions: 1) Century 2) Age, Era, Period 3) Stage, Phase, Subperiod 4) Ambiguous temporal expression 5) Cyclical temporal expression.

We have implemented above five cases with enhancements of ISO 19108 specifications, as follows. (In following cases, class names that start with "TM_" are from ISO 19108)

For case 1): It is the common use to specify a century number as the temporal expressions, like as "8th century". Since ISO 19108 does not support the reference of the century number, we have defined "Common Century System" class for century orders as a temporal reference system. This class is inherited from the TM_Calendar class. And, to express a specific century number, we have also defined "Common Century" class which is inherited from the TM_coordinate class.

For case 2): It is also the common use to specify Age/Era/Period name as the temporal expressions, like as “Kamakura period (鎌倉時代)”. ISO 19108 defines the ordinal reference system and its element, but it is not fit for the practical use of Age/Era/Period name as the temporal expressions for historical or archaeological objects. We have defined “Chronological Reference System” class to identify the chronological order. This class is inherited from the TM_OrdinalRefenceSystem. And we have defined “Chronological Element” class, that is inherited from the TM_OrdinalEra class, to express each age/era/period names.

Case 3) includes the expressions, e.g., “early stage (前期)”, “the beginning (初頭)”, “the first half (前半)”. This type of qualification defines the part in the range of original period. It is not defined in ISO 19108. We have implemented it by adding “periodical qualifier” attribute in the class definition for the period.

Case 4) includes the expressions, e.g., “from the end of 7th century to the beginning of 8th century (7 世紀末から 8 世紀初頭)”, “from the last stage of Nara Period to the beginning of Heian period (奈良時代後期から平安時代初頭)”. We have defined a class that accepts two or more types of the instance including case 3) with optional attribute of the estimated probability.

Case 5) includes the expressions, e.g., “kanoto-i year in Kofun period (古墳時代の辛亥の年)”, “winter in the latter portion of Meiji period (明治時代後葉の冬)”. In these examples, “kanoto-i” is a 48th year in Jikkan (The Ten Stems: 十干) and Junishi (the Twelve Signs of the Chinese Zodiac: 十二支) in the period for 60 years cycle, and “winter” is one of the four seasons in a year cycle. We have added the function expression at the “periodical qualifier” attribute in the class definition of period. The rectangular wave function is a practical case for implementing cyclical temporal expressions.

Our approach to enhancing ISO 19108 will be possible to lead the standardization of the temporal expressions for historical or archeological objects on information system. It also will provide the common temporal specification not only for the history or archaeology such as the studies treating the past, but also for the whole field of humanities.

Keywords
temporal attribute, history, archaeology, chronology

The Echo of Print: Outing Shakespeare's Source Code at St Paul's

Thomas W Dabbs (Aoyama Gakuin University)

This talk will examine how digital platforms in development may be used to undo a scholarly dogma that has historically restricted our understanding of Shakespearean drama. Traditionally these dramas have been viewed as privileged primary literature that has been fused with lesser secondary sources by a singular creative genius. The use of the term *source* suggests to us that plot of *Romeo and Juliet*, for instance, was drawn from minor or obscure print editions that the Bard of Avon molded into a fine literary work.

This line of reasoning is flawed. To view Ovid's *Metamorphoses* and its popular Elizabethan translation into English by Arthur Golding as secondary to Shakespeare's frequent use of this edition, is comparable to saying that J. R. R. Tolkien's *Lord of the Rings* is secondary to the film adaptations of the same story. Many of the so-call sources that Shakespeare and other playwrights used, for instance the popular collection of stories in William Painter's *Palace of Pleasure*, were more prominent in the minds of the Elizabethan public than the plays adapted from them. By cross-referencing searchable databases and digital reconstructions of Elizabethan London, we can see that Shakespearean drama was in fact keenly adapted to the popular reception of printed works available in English, particularly in the St Paul's precinct.

This talk will examine digital reconstructions of the St Paul's cathedral precinct in the City of London, the center of the book selling industry during the Elizabethan period. The cathedral's great and boisterous nave, Paul's Walk, and the open churchyard full of bookshops at Paul's Cross, were centers for broadcasting new print.

Until recently, however, it has been difficult to visualize this enormous locale as it existed during the Elizabethan period. Digital reconstructions of the cathedral precinct show that Shakespearean plays and many other plays were not crafted from obscure or lesser books. Instead such plays echoed from local theatres the reception of popular printed works particularly in the public sphere at St Paul's.

As a work sample, this talk will examine the single example of William Painter's *Palace of Pleasure*. This popular work comprised Painter's translations of many classical and continental stories, including, among other Shakespearean adaptations, the stories of Romeo and Juliet and Timon of Athens. This publication was used by pre-Shakespearean playwrights to craft a spate of plays after its popular reception in the City of London and specifically at St Paul's. By the time Shakespearean plays reached the public stage, the use of Painter and other popular authors had become something of a template for staging successful productions.

Figure 1. Cropped from a 17th-century Dutch painting (Museum of London), showing the enormity of St Paul's and its proximity with the public theatres flying their flags.

Several digital initiatives will be used to show the progress from the printing of Painter's work to its open public reception with stories from it being adapted for the Elizabethan stage, including adaptations by Shakespeare. The talk will begin with the *Agas Map of London* online in order to show how St Paul's was positioned in the City of London in relation to local theatres that came into being within and on the outskirts of the city. The reconstructions at the *Virtual Paul's Cross Project*, will show the physical environment of the cathedral proper and also a reconstruction of Peter Blayney's (hard copy) map of the bookstores of Paul's Cross churchyard. These reconstructions point to the fact that new printed works were often read and discussed in this locale. Such databases

as *EEBO-TCP* and the *ESTC* online will also be used to confirm the popular reception of Painter's work within the St Paul's precinct.
Titles of extant plays will be used with titles in the *Lost Plays Database* to show how early modern plays were crafted, not from singular inspirations drawn from independently selected source material, but from playwrights, including Shakespeare, hearing the echoes of popular printed works specifically in the St Paul's precinct. The relationship between popular stories and plays can be established by searching *EEBO-TCP*, the *ESTC*, and other online reference material and then cross-referencing stories with play titles. The presumed story in the lost play, 'Cupid and Psyche' will not show up in a reading of Painter's table of contents, but 'A Greek Maid' will, if one recognizes that the story of 'Timoclea of Thebes' is indeed about a Greek maiden and is the probable source of 'Greek Maid'. The methodology here is much easier to show in *Powerpoint* than to describe in abstract, but the base method is to fill a reconstructed public gathering site with bookshops and popular stories that echoed into successful stage plays during the early modern period. The talk will conclude that such stories as those found in Painter are not source material, per se, but well-known stories that were read and discussed in a central bookselling area and that were later cherry picked because of their apparent popular appeal to be adapted for commercial theatre events.
Along with showing how DH platforms can be collaborated, three suggestions will be made for the future of early modern digital development and scholarship. The first concerns the singular direction of DH projects and the current need to increase the interoperability between platforms. For instance the *Virtual Paul's Cross Project* recreates the environment at Paul's Cross churchyard to focus on a sermon by John Donne. It is not currently aimed to provide more information about the churchyard bookstores that the project accurately reconstructs or information about printed editions on sale in these bookstores. This problem could be solved with the inclusion of an interactive interface that would provide pop-up bubbles with information about churchyard bookshop holdings. These bookshop holdings could in turned be linked to full texts (when available) at *EEBO* and to publication information at the *ESTC*.
The second suggestion concerns the unfinished nature of these projects. *EEBO-TCP* is slow in development as are other projects. This subject will be mentioned only is passing as it could be the focus of an entire DH conference, one that would focus on how to manage continuous and reliable data input for open access sites.
The third suggestion is rooted in the fact that some of our greatest resources are only preserved in hard copy, with no search-ability at all or just the 'look inside' option at Amazon or the frustratingly narrowed options offered by Google Books. The future for digital research in the early modern period is in seeing ways to continue the development and interoperability of existing databases with interactive interfaces. We should find ways to finish and better collectivize what has been started, and to digitalize information in hard copy texts in ways more elegant than simple reproductions of the text.

References

Primary Texts (Modern spelling)
[1] Bower, Richard? *Apius and Virginia* (London: Richard Jones, 1575). Full text: *EEBO-TCP*. Gosson, Stephen. *Plays Confuted in Five Actions* (London: Thomas Gosson, 1582). Full text: *EEBO-TCP*.
[2] Naso, *Ovid. The XV Books of P. Ouidius Naso, entitled Metamorphosis*. Trans. Arthur Golding (London: William Seres, 1567). Full text: *EEBO-TCP*.
[3] Painter, William. *The Palace of Pleasure* (London: Richard Tottell, 1566). Full text: *EEBO-TCP*.
[4] Shakespeare, William. *The Most Excellent and Lamentable Tragedy of Romeo and Juliet* (London: Cuthbert Burby, 1599). Full text: *Internet Shakespeare Editions*.
[5] Wilmot, Robert? *The Tragedy of Tancred and Gismund* (London: R. Robinson, 1591). Full text: *EEBO-TCP*.

Lost Plays (Bibliographic Entries)
[6] From *Lost Plays Database*. Ed. Roslyn L. Knutson and David McInnis (Melbourne: University of Melbourn, 2009).
Anon. 'A Greek Maid' (1579). Thomas Dabbs. Web. https://www.lostplays.org/index.php?title=Greek_Maid,_A.

Anon. 'A Mask of Amazons' (1579). (Forthcoming. See Wiggins below.)
Anon. 'Mutius Scaevola' (1577). Thomas Dabbs (forthcoming).
Anon. 'The Story of Samson' (1576). Roslyn L. Knutson. Web. https://www.lostplays.org/index.php?title=Samson.
Anon. 'Timoclea of Thebes' (1574). John H. Astington. Web. https://www.lostplays.org/index.php?title=Timoclea_at_the_Siege_of_Thebes.

Digital Projects
[7] *Digital Renaissance Editions*. Web. http://digitalrenaissance.uvic.ca.
[8] *Internet Shakespeare Editions*. Web. http://internetshakespeare.uvic.ca
[9] *Lost Plays Database*. Web. https://www.lostplays.org/index.php?title=Main_Page.
[10] *Map of Early Modern London (MoEML)*. Web. https://mapoflondon.uvic.ca.
[11] *Shakeosphere*. Web. https://shakeosphere.lib.uiowa.edu.
[12] *Shakespeare Quartos Archive*. Web. http://www.quartos.org/index.html Stow, John, *A Survey of London*: *From the Text of 1603* in *(BHO)*. Web. http://www.british-history.ac.uk/no-series/survey-of-london-stow/1603.
[13] *The Virtual Paul's Cross Project*. Web. https://vpcp.chass.ncsu.edu.

Databases
[14] *Database of Early English Playbooks (DEEP)*. Web. http://deep.sas.upenn.edu.
[15] *Early English Books Online (EEBO-TCP)*. Web. http://quod.lib.umich.edu/e/eebogroup.
[16] *English Short Title Catalogue (ESTC)*. Web. http://estc.bl.uk.
[17] *Hamnet: Folger Library Catalog*. Web. http://shakespeare.folger.edu.
[18] *Records of Early English Drama (REED)*. Web. http://reed.utoronto.ca.

Workshop Resources
[19] *Early Modern Digital Humanities: Japan (EMDH: Japan)*. 'Master List of Resources.' comp. John Yamamoto-Wilson Web. http://emdhjapan.blogspot.jp/2014/03/dh-database- links.html.

Hard Copy (limited digital search, Google)
[20] Dabbs, Thōmas. 'Paul's Cross and the Dramatic Echoes of Early-Elizabethan Print' in *Paul's Cross and the Culture of Persuasion in England, 1520-1640*. Ed. Torrance Kirby and P. G. Stanwood (Leiden: Brill, 2014).
[21] Gurr, Andrew. *Playgoing in Shakespeare's London* (Cambridge: Cambridge UP, 1987; rpt. 2004). Morrissey, Mary. *Politics and the Paul's Cross Sermons, 1558-1642* (Oxford: Oxford UP, 2011). Shakespeare, William, *Romeo and Juliet* ed. René Weis (London: Arden, 2012).
[22] Wiggins, Martin. *British Drama 1533-1642: A Catalogue*. Vol. II and Vol. III (Oxford: Oxford UP, 2012).

Hard Copy Only (providing some scanned images)
[23] Blayney, Pēter M.W. *The Bookshops in Paul's Cross Churchyard*. (London: The Bibliographical Society, 1990).
[24] MacLure, Millar. *The Paul's Cross Sermons* (Toronto: University of Toronto Press, 1958).
[25] Schofield, John. *St Paul's Cathedral before Wren* (Swindon: English Heritage, 2011). *St Paul's. The Cathedral Church of London:604-2004*. Ed. Derek Keene, Arthur Burns, and Andrew Saint (New Haven: Yale UP, 2004).

Comparing Topic Model Stability across Language and Size

Simon Hengchen (Université libre de Bruxelles), Alexander O'Connor (ADAPT Centre School of Computing, Dublin City University), Gary Munnelly (ADAPT Centre, Trinity College Dublin), Jennifer Edmond (Long Room Hub, Trinity College Dublin)

The rapid evolution of technology has freed the written word from the physical page. In the current era, it can be argued that the primary means of access to text is digitally mediated. This has given unprecedented reach to any individual with access to the Internet. However, the rate at which a human can absorb such information remains relatively unchanged, in particular in the case of linguistically and/or culturally complex data. Results in computer science continue to advance in areas of linguistic analysis and natural language processing, facilitating more complex numerical inquiries of language. This commoditisation of analytical tools has led to widespread experimentation with digital tools within the humanities: recent initiatives such as DARIAH[1], CENDARI[2] or TIC-Belgium[3] try to foster the use of computational methods and the reuse of digital data by and between researchers and practitioners alike.

A key question emerges: to what extent do these digital tools reveal signal, and to what extent are they merely responding to noise? This is a question of particular import to human- ities researchers, for whom the difference between signal and noise may shift from project to project and from interpreter to interpreter, not to mention from linguistic context to linguistic context. Scholars currently must resort to a vehicular language (in Europe and North Amer- ica, generally English) in order to find patterns between cultural and linguistic contexts. This approach is not wholly satisfying, however, where the sensitivities surrounding the object of study are high, meaning that speakers would choose specific words and phrases with great care, aware of the resonances of the choices.

Discourse regarding cultural traumas, such as war, occupation, economic collapse, envi- ronmental disaster, or other major disruption to national identity and social cohesion, present a clear example of this kind of issue: culturally specific, and yet present at some level or other in nearly every cultural narrative. The international SPECTRESS network[4] had hoped to provide
a new approach to fostering cross-cultural dialogue regarding the impact of and responses to cultural trauma by topic modelling discourse around traumas, and seeking similar clustering effect across language- and event-specific contexts. The challenge with this approach was that appropriate corpora were generally too small to produce reliable models and results. However, initial experiments were not able to answer one key question of interest to both the computer scientists and the humanists in the project team: how small is too small?

We focus on the study of language and the semi-automatic discovery of topics in textual data. In order to extract meaning we use two algorithms, both often referred to as "topic mod- elling techniques": Latent Semantic Analysis (LSA) (Landauer et al., 1998) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003). Both algorithms construct matrices to try to determine topics within a set of texts by clustering similar words. These approaches both encode key assumptions about the statistical properties of the language, with statistical and stochastic as- pects included. Whilst LDA is the most widely used algorithm in the literature these past years, we believe that a benchmarking study should include more than one take at the data, which is why we are comparing LDA and LSA. Both models also need a certain pointed out by Greene *et al* (Greene et al., 2014). Unfortunately, it is unclear how much data is enough. This lack of clear understanding of minimal functional corpus size poses a serious threat to topic modelling's viability as humanistic methodology. Topic modelling is currently an approach humanists are very aware of and see potential uses for (following the work of Jockers (Jockers, 2013; Jock- ers and Mimno, 2013) and others), but as many humanistic corpora are on the small side, the threshold for the utility of topic

1 http://dariah.eu/
2 http://cendari.eu/
3 http://tic.ugent.be/

modelling across DH projects is as yet highly unclear. Unsta- ble topics may lead to research being based on incorrect foundational assumptions regarding the presence or clustering of conceptual fields on a body of work or source material. Stable topics, however, indicate that the random component in the process has been minimised and the topics given do possess a coherence worthy of further investigation by a trained human, as advocated by Chang *et al* (Chang et al., 2009).
Building on previous work by Munnelly *et al* (Munnelly et al., 2015), we propose a method- ology to try to determine how large a corpus must be to establish a stable model, with an added twist: whilst topic modelling techniques are language-independent, i.e. "use[] no manually- constructed dictionaries, knowledge bases, semantic networks, grammars, syntactic parsers, or morphologies, or the like."(Landauer et al., 1998), the morphology of the language processed can influence the size of the corpus required to build a stable set of topics. In order to do so, we compare French and English topic models from a bilingual corpus of articles.

Methodology

We use the DBpedia (Auer et al., 2007) interlanguage links for the English lan- guage (interlanguage-links_en.nt) to search for every DBpedia URI existing in French and in English[5].
With all DBpedia URIs having a match – and linked via the owl:sameAs predicate – in both languages, we then parse both long_abstracts_en.ttl and long_abstracts_fr.ttl files to extract their respective long abstracts.
This process carried through, we decompose the resulting files in a number of smaller files: one for every DBpedia entity, each containing its abstract. With both corpus segments consti- tuted, it is possible to apply LSA and LDA. The resulting models are stored and measured. The corpora are reduced in size, LDA and LSA re-applied, models stored, and corpora re-reduced, iteratively, each time recording the topic results.
Topic models are compared manually between languages at each stage, and programmati- cally between stages, using the Jaccard Index (Real and Vargas, 1996), for both languages.
A large deviation between stages indicates a loss of representativeness between models.

Perspectives

By applying our methodology on parallel corpora, we try to determine whether the minimum sample size for a representative topic model is consistent across the two lan- guages studied, i.e. French and English. Using the built-in multilingualism of DBpedia, it be- comes possible to reapply the methodology on most written languages.

References

[1] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). *Dbpedia: A nucleus for a web of open data*. Springer.
[2] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.
[3] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296.
[4] Greene, D., O'Callaghan, D., and Cunningham, P. (2014). How many topics? Stability analysis for topic models. In *Machine Learning and Knowledge Discovery in Databases*, pages 498–513. Springer.
[5] Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press.
[6] Jockers, M. L. and Mimno, D. (2013). Significant themes in 19th-century literature. *Poetics*, 41(6):750–769.
[7] Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

4https://spectressnetwork.wordpress.com/
5The files are freely available for download at http://wiki.dbpedia.org/Downloads2015-10.

[8] Munnelly, G., O'Connor, A., Edmond, J., and Lawless, S. (2015). Finding meaning in the chaos.
[9] Real, R. and Vargas, J. M. (1996). The probabilistic basis of jaccard's index of similarity. *Systematic biology*, 45(3):380–385.

Can a writer disguise the true identity under pseudonyms?: Statistical authorship attribution and the evaluation of variables

Miki Kimura (Meiji University)

This is a work-in-progress study on quantitative authorship attribution of a lesbian writer with more than one pseudonyms, James Tiptree, Jr. and Raccoona Sheldon. Alice Bradley Sheldon (1915-1987) was a writer who published feminist science fiction stories for almost 20 years. As a commercial strategy, she hid her true identitiy under a male pseudonym, James Tiptree, Jr., for little over a decade. She also used a female pseudonym, Raccoona Sheldon, as the name offered a thematic change.

Brinegar (1963) inspected the distribution of word length in order to verify the author of the QCS letters and concluded that the letters were not written by Mark Twain. Mosterller and Wallace's study of the Federalist papers verified the author of a collection of eighteenth-century political documents, which argue for the Constitution of the United States, through the frequencies of individual words such as prepositions, which are considered irrelevant to the content of the papers. Burrows (1987) examined intra-author variations in Jane Austen's novels by employing a statistical method called principal component analysis. In Japan as well, stylometry has developed over the past 50-plus years. In particular, Jin, Kabashima, and Murakami (1993) inspected intra-author variation in the works of a well-known Japanese author who used three pseudonyms. They could not detect intra-author variation in the Japanese author's works, but they were able to show inter-author variation in comparison with the author's contemporaries by using the distribution of commas in Japanese.

In this research, I will examine intra/inter author variations in Alice Sheldon's texts. As Le Guin (1976) indicated Alice Sheldon's works under the female pseudonym (Raccoona Sheldon) have less control and wit compared to her works under the male pseudonym (James Tiptree, Jr.). Using statistical analyses, this research primarily focused on the intra-author variation between her works under these two pseudonyms. It not only distinguished Alice Sheldon's works under the two pseudonyms but also compares the results from this quantitative authorship attribution with the works of literary criticism scholars such as Silverberg (1975), Lefanu (1989), Russ (1995), and Larbalestier (2002).

In addition to the examination of intra-author variasion within the works of one author, this research also investigates inter-author variation between two authors. As Silverberg (1975), Lefanu (1989), and Kotani (1994) noted, in contrast to Ernest Hemingway, James Tiptree's manner of writing is somewhat masculine. In order to address such criticisms, the Alice Sheldon's Corpus, which consists if all the works publish under her two pseudonyms, and the Hemingway Corpus, which contains all his short stories, have been developed.

Juola (2013) recently inspected intra-author variation in the works of Joanne Rowling, who uses the two pseudonyms J. K. Rowling and Robert Galbraith, and tries to attribute the works under Robert Galbraith to those written under J. K. Rowling. This study used a specialized software called JGAAP, and verified that the works under J. K. Rowling and those under Robert Galbraith have the same style as other British female writers. Further, according to a case study on the quantitative stylistics of Joanne Rowling presented by Kimura and Kubota (2015), the author skillfully differentiates her writing style by genres and pseudonyms.

This result could possibly be useful for the analysis in the current study. However, another probable assumption is that author discriminators chosen form the corpora developed for this kind of research differentiate between the two authors, but fail to discriminate between Alice Sheldon's two pseudonyms. The latter result means that Alice Sheldon failed to disguise her true identity by using the two pseudonyms James Tiptree, Jr. and Raccoona Sheldon.

As variables, the top 10, 25, 50 most common words, considered effective for this kind of discrimination by, for example, Burrows and Hassall (1988) and Burrows (1992), are chosen for the analysis. In addition to these lexical variables, this research has also selected syntactic variables, especially the distribution of POS, which are considered effective for discrimination based on Hirst and Feiguina (2007). I will apply two kinds of unsupervised statistical methods

(principal component analysis and hierarchical clustering analysis) and two supervised classification methods (discriminant analysis and support vector machines — SVM). If the discrimination variables chosen from these two corpora have sensitivity as identifiers, the results from SVM will show that they can capture inter-author variation between works from Alice Sheldon and works from Ernest Hemingway, but cannot detect intra-author variation between works under Alice Sheldon's two pseudonyms. In this analysis, the evaluation of the classification methods and the variables, which are considered effective for such research, will be simultaneously conducted.

References

[1] Burrows, J. F. (1987) Computation into Criticism: A study of Jane Austen's novels and an experiment in method. Oxford: Clarendon Press.

[2] Burrows, J. F. (1992). Not unless you ask nicely: The interpretative nexus between analysis and information. Literary and Linguistic Computing, 7(2), 91-109.

[3] Burrows, J. F., & Hassal, A. J. (1988). Anna Boleyn and the authenticity of Fielding's feminine narratives. Eighteenth Century Studies, 21, 427-453.

[4] Hirst, G. & Feiguina, O. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. Literary and Linguistic Computing, 22(4), 405–417.

[5] Russ, J. (1995). To write Like a Woman. Bloomington: Indiana University Press.

[6] Silverberg, R. (1975). Who Is Tiptree, What Is He? Warm Worlds and Otherwise. New York, Ballantine Books. iv - x viii

[7] 金明哲・樺島忠夫・村上征勝 (1993). 「読点と書き手の個性」 『計量国語学』 18(8), 382–391.

[8] Juola, P. (2013, July 16). Language Log: Rowling and "Galbraith": an authorial analysis. Retrieved from http://languagelog.ldc.upenn.edu/nll/?p=5315

[9]木村美紀・久保田俊彦 (2015). 「男女両名義を使用する作家の作品判別― Rowling と Sheldon」, 第 41 回英語コーパス学会発表資料.

[10]小谷真理 (1994). 『女性状無意識: テクノガイネーシス―女性 SF 論序説』 東京: 勁草書房, 40-67

[11] Larbalestier, J. (2002). The Battle of the Sexes in Science Fiction. Connecticut: Wesleyan University Press.

[12] Lefanu, S. (1989). Who Is Tiptree, What Is She? : James Tiptree, Jr.. Feminism and Science Fiction. Bloomington: Indiana University Press.

[13] Le Guin, U. K. (1978). Introduction. Star Songs of an Old Primate. New York: Ballantine Books. vii - x ii

[14] Mosteller, F., & Wallace, D. L. (1964). Inference and Disputed Authorship: The Federalist. Reading, MA: Addison-Wesley.

Associative Network Visualization and Analysis as a Tool for Understanding Time and Space Concepts in Japanese

Maria Telegina (University of Oxford)

The history of graph (network) theory (GNT) started with an attempt to find a single walking path, which crosses, once and only once, each of the seven bridges of old Königsberg; this is known as the Seven Bridges of Königsberg Problem. Since 1736, when Leonhard Euler proved the problem to be unsolvable using a very simple graph, GNT was developed, and it rapidly come to be used in a number of fields. Nowadays, GNT is actively used in a wide variety of disciplines from mathematics and physics to sociology and linguistics (e.g., Mehler, A., et al, 2016), as our world is full of systems, which can be represented and analyzed as networks.

The main focus of this paper is a presentation of a network visualization and analysis, based on an association network constructed on Japanese temporal and spatial lexical items. The network (Fig.1) is based on the results of an ongoing free word association experiment, the first stage of which was conducted in Tokyo in 2015, involving 85 native Japanese speaking participants of two different age groups (one in their 20s and one from their 50s to 70s).

Particular temporal and spatial lexical items for the experiment were selected on the basis of four main sources: A Frequency Dictionary of Japanese (2013), Japanese Word Association Database ver. 1 (2004), Associative Concept Dictionary (2004, 2005) and Japanese WordNet ver. 1.1. The criteria for the selection were based on a variety of frequencies according to Frequency Dictionary of Japanese (from *Toki* with 2514 occurrences per million words to *Ima* with 9 occurrences per million words) and a variety of semantic relations within the stimuli set (synonyms, hyponyms, antonyms).

Synonyms (partial synonyms) are represented by *kuukan, supeesu, yochi, hirogari; basho, ba; sukima, suki; ima, ribingu; aida, ma; jikan, toki, taimingu*; *kyuujitsu, yasumi, hima; basho, ba; wagaya, mai hoomu; nagasa, kyori; hizuke, hi.* Synonyms are chosen in accordance with WordNet.

The hyponyms and hypernyms in this study are *heya, apaato, manshon/ie; aki, natsu/ kisetsu; jidai, jiki, naganen, kisetsu, shunkan, hi/jikan; asa, yoru, hiruma/hi, oku, ie /kuukan*; *mukashi/toki.* Hyponyms and hypernyms are selected in accordance with the Japanese Word Association Database and the Associative Concept Dictionary.

Also, *soto, uchi; mae, ushiro; kako, mirai; tonai, kougai* were selected as antonyms or opposites in accordance with the Japanese Word Association Database and the Associative Concept Dictionary.

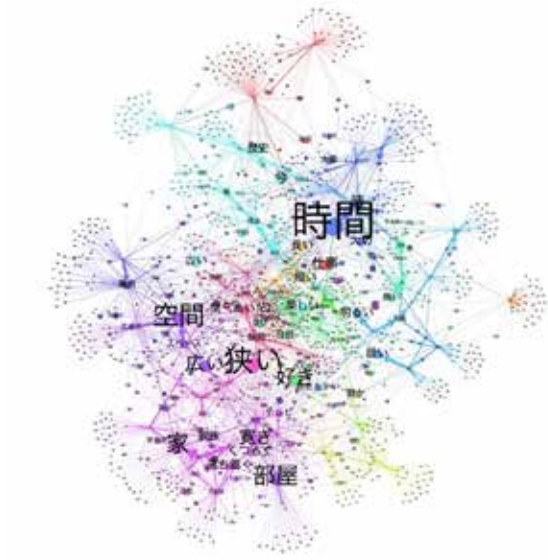

Figure 1.

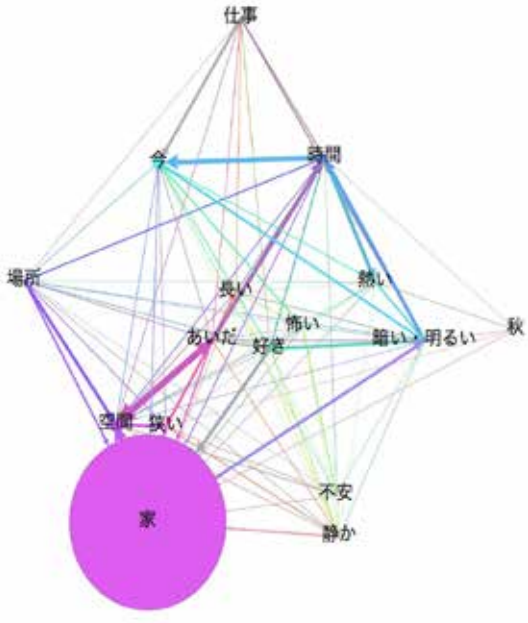

Figure 2.

Ten fillers were chosen randomly with the criterion to cover approximately the same frequency range as within the stimuli set. The fillers were added to the survey to serve as distraction from temporal and spatial stimuli words and to inimize the number of deliberate responses; the responses to the fillers are not included in the analysis.

The main purposes of this study are on three different levels: first, a macro-level, discussing the possibility of utilizing the association network analysis to describe the conceptual structure of the language in question; second, a meso-level, analyzing communities formed within the network; and third, a micro-level, investigating the usage of association networks to formulate the cognitive definitions of single words within the network by identifying their features based on their connections within the network.

At this stage of analysis, the findings suggest that the analysis of single word connections and their weight might be utilized for disambiguation of meanings of synonymic words for cognitive definitions (Ostermann, C., 2015). It demonstrates information which could be also found in traditional dictionary definitions or corpora materials, such as typical syntagmatic connections, e.g., *sukima-kaze* and *suki-yudan.* At the same time, culturally specific semantic features of the lexical items, which can hardly be predicted through the materials based on the common language production, e.g., *ushiro-kowai* or both negative and positive emotional evaluation of *hima,* can be found.

At the meso-level, ten communities, e.g., abstract space, concrete (physical) space, life time, dark/light time, home, etc., were detected within the network using the Order Statistics Local Optimization Method. The stru cture of connections between the communities is complex with numerous overlaps. However, on the basis of the inter-communities connections, it is still possible to hypothesize about a macro-level conceptual structure of Japanese, e.g. based on this analysis, it could be concluded that temporal and spatial concepts in modern Japanese are the most closely connected to two concepts: emotional evaluation and daily life (Fig. 2).

Finally, on the basis of this analysis, I propose an associative network as an illustrative and effective tool for planning further experimental work.

References

[1] Caldarelli, G. (2007). Scale-free networks: Complex webs in nature and technology. Oxford: Oxford: Oxford University Press.

[2] Dorogovtsev, S. N. (2010). Lectures on complex networks. Oxford: Oxford: Oxford University Press.

[3] Japanese Wordnet (v1.1), copyright NICT, 2009-2010 Joyce, T. Large-scale Database of Japanese Word Associations, Version1, http://www.valdes.titech.ac.jp/~terry/jwad.html

[4] Lancichinetti, A., Radicchi, F., Ramasco, J.J., Fortunato S. (2011). Finding statistically significant communities in networks. PLoS ONE 6: e18961.

[5] Mehler, A., Lücking, A., Banisch, S., Blanchard, P., & Job, B. (Eds.). (2016). Towards a theoretical framework for analyzing complex linguistic networks. Berlin: Springer Berlin Heidelberg.

[6] Newman, M. E. J. (2010). Networks: An introduction. Oxford: Oxford: Oxford University Press.

[7] Okamoto, J., Ishizaki, S. (2004, 2005) Rensoogainenjisho. Associative Concept Dictionary Ostermann, C. (2015) Cognitive Lexicography. A New Approach to Lexicography Making Use of Cognitive Semantics, Berlin, Boston: De Gruyter Mouton

[8] Tono, Y., Yamazaki M., Maekawa K. (Eds.). (2013). A frequency dictionary of Japanese: Core vocabulary for learners. London : Routledge

Melodic Structure Analysis of Traditional Japanese Folk Songs from Shikoku District

Akihiro Kawase (Doshisha University)

Introduction

This study aims to grasp the regional differences in the musical characteristics inherent in the traditional Japanese folk songs by extracting and comparing the characteristics of each area by conducting quantitative analysis in order to promote digital humanities research on traditional Japanese folk songs.

In the previous studies, We have sampled 1,794 song pieces from 45 Japanese prefectures, and have clarified the following three points by extracting and comparing their respective musical patterns (Kawase and Tokosumi 2011): (1) the most important characteristics in the melody of Japanese folk songs is the transition pattern, which is based on an interval of perfect fourth pitch; (2) regionally adjacent areas tend to have similar musical characteristics; and (3) the differences in the musical characteristics almost match the East-West division in the geolinguistics or in the folkloristics from a broader perspective. However, to conduct more detailed analysis in order to empirically clarify the structures by which music has spread and changed in traditional settlements, it is necessary to expand the data and do comparisons based on the old Japanese provinces (ancient administrative units that were used under the ritsuryo system before the modern prefecture system was established).

In this study, we analyzed all the songs listed from the Shikoku district (literally meaning four provinces, located south of Honshu and east of Kyushu district) in order to build a digital analysis platform for all the songs recorded in the *Nihon Min'yo Taikan* (Anthology of Japanese Folk Songs) and execute quantitative comparisons of musical characteristics between neighboring regions (Kawase 2016a; 2016b).

Procedure

Specifically, the procedures are as follows: (1) we digitized all the songs from the Shikoku district and generated sequences that contain interval information from the song melodies;
(2) extracted patterns that appear with high frequency in the generated sequences; and (3) summarized the musical characteristics of the folk songs from the Shikoku district by comparing the patterns between provinces using statistical techniques.

In order to digitize the Japanese folk song pieces, we generate a sequence of notes by converting the music score into MusicXML file format. We devised a method of digitizing each note in terms of its relative pitch by subtracting the next pitch height for a given MusicXML. It is possible to generate a sequence T that carries information about the pitch to the next note: $T = (t1, t2, \ldots, ti, \ldots, tn)$. An example of the corresponding pitch intervals for ti can be written as shown in **Table 1**. We treat sequence T as a categorical time series, and execute N-gram analysis by conducting unigram, bigram, and trigram patterns to clarify major transitions and their trends in the Shikoku district.

Results

Based on the results of N-gram analysis, we found that folk songs from the Shikoku district have a strong tendency to form melodic leaps followed by progressions back to the first sung note or perfect fourth intervals, as a characteristic of N=1, 2, 3 interval transition pattern. In particular, patterns where the total of the elements themselves for N=2 form perfect fourth intervals are the ascending and descending order for the four types of tetrachords that Fumio Koizumi proposed (Koizumi 1958). In addition, patterns that include N=3 tetrachords also were extracted remarkably

Table 1: Corresponding Pitch Intervals

t_i	Pitch Intervals	t_i	Pitch Intervals
0	perfect unison	7	perfect fifth
1	minor second	8	minor sixth
2	major second	9	major sixth
3	minor third	10	minor seventh
4	major third	11	major seventh
5	perfect fourth	12	perfect octave
6	aug.fourth/dim.fifth	13	minor ninth

often.

The tetrachord is a unit consisting of two stable outlining tones with the interval of a perfect fourth pitch, and one unstable intermediate tone located between them. Depending on the position of the intermediate tone, four different types of tetrachords can be formed (**Table2**). Below are some discussions about the features of folk songs, focusing on interval transitions that form tetrachords.

Table 2: For Basic Types of Tetrachords

Type	Name	Pitch Interval
I	*Min'yo*	minor third (3) + major second (2)
II	*Miyako bushi*	minor second (1) + major third (4)
III	*Ritsu*	major second (2) + minor third (3)
IV	*Ryukyu*	major third (4) + minor second (1)

Discussion

Out of four types, we found that *min'yo* tetrachords were used with an extremely high frequency, and the next highest was *ritsu tetrachords*. Furthermore, we conducted a cluster analysis (hierarchical clustering) based on the frequency of occurrences of the tetrachords to see the differences in each province (see **Figure 1**). When calculating distances between each element, we normalized the frequency that the tetrachords appear, and used the Euclidean distance and the algorithm from the Ward method.

Compared with our previous analysis on neighboring regions such as the Kyushu and Chugoku districts (Kawase 2015; 2016ab), we find that folk songs from the eastern two provinces (Sanuki and Awa) and western two provinces (Iyo and Tosa) of Shikoku district can be explained in terms of differences in melodic structures within tetrachords. In particular, for western provinces, there is a tendency to create the ritsu and ryukyu tetrachords, which also appear frequently in Kyushu district. In contrast, for eastern provinces, there is a tendency to create the miyakobushi tetrachord, which is thought to be originated from music of urban areas such as in Kyoto. Thus, the tetrachord turned out to be salient characteristic by which to classify the melodies of east and west regions of Shikoku district.

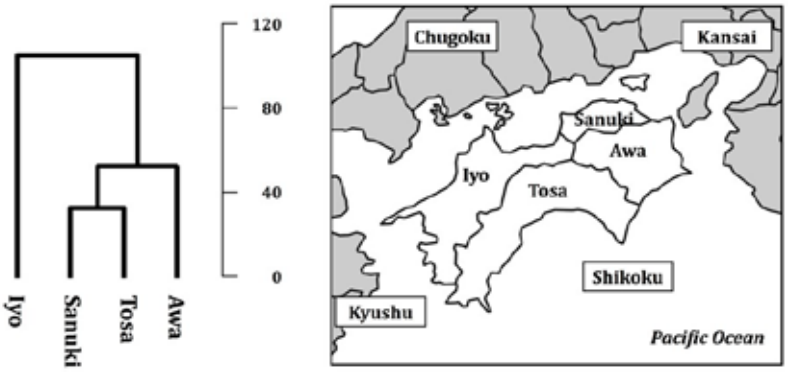

Figure 1: Dendrogram based on transition probabilities of tetrachords for four provinces

Acknowledgements

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

References

[1] Kawase, A. (2016a) Regional classification of traditional Japanese folk songs from the Chugoku district, In *Proceedings of the Digital Humanities 2016: DH2016* (in press).

[2] Kawase, A. (2016b) Extracting the musical schemas of traditional Japanese folk songs from Kyushu district, In *Proceedings of the 14th International Conference for Music Perception and Cognition: ICMPC14* (in press).

[3] Kawase, A. and Tokosumi, A. (2011) Regional classification of traditional Japanese folk songs, *International Journal of Affective Engineering* 10 (1): 19-27.

[4] Koizumi, F. (1958) *Nihon dento ongaku no kenkyu (Studies on Traditional Music of Japan 1)*, Ongaku no tomosha.

[5] Nihon Hoso Kyokai (1944-1993) *Nihon Min'yo Taikan (Anthology of Japanese Folk Songs)*, Nihon Hoso Kyokai Shuppan.

[6] MusicXML http://www.musicxml.com/for-developers/ [accessed 15 May 2016].

Visualizing Japanese Culture Through Pre-Modern Japanese Book Collections—A Computational and Visualization Approach to Temporal Data—

Goki Miyakita, Keiko Okawa (Keio University)

This paper proposes a design of online digital collection of pre-modern Japanese books by using computational and visualization approach to open a new vision of Japanese culture through books. Digital collections as an emerging field has made significant changes in the way we interact with books from physical to virtual, however, most collections places their emphasis on only digitization and academic uses, and focuses less on its visualization and use by the general public. Therefore, the aim of this research is to explore historical temporal data, namely rare Japanese books from the 8th to the 19th century with advanced computer-based visualization approach, and to reveal the cultural history, trends, and fashion in Japan in narrative form. This research will examine the method of digitization and visualization in a coherent manner, in order to enable diverse audience to access, browse, and interact with the vast collection from Keio University's collection of pre-modern Japanese books.

During the past few years, there is a dramatic shift in the way we preserve books. This shift allows books to exist not only as a genuine artifact but also as a replicated or restructured digital artifact that exists in the virtual world. Ever since the emergence of the Internet and the World Wide Web, printed books—especially, the books that is distinguished by its early printing date, namely rare and pre-modern book collections—has transformed an ontology from physical to the virtual space by offering the promise of new forms and content delivery that exceed the limitations of printed. However, most researches in Japan remains in developing their digitization techniques, creating a database or an online-archive for academic usage. Therefore, it is difficult for the general audiences—especially for those who does not understand Japanese or does not possesses knowledge related to Japan—to improve their understanding of Japanese culture through pre-modern Japanese books.

The research presented in this paper proposes a new conception of digital collection through practice-led research. I work with the collection from Keio University's Institute of Oriental Classics, which keeps extensive collection specialized in pre-modern Japanese books from the 8th to the 19th century, combine and adapt computational and visualization approaches to interpret information of the books and to promote understanding of Japanese culture for the audiences from a wide range of nationalities and backgrounds. Furthermore, the digital collections are implemented to a Massive Open Online Course (MOOC): *Japanese Culture Through Rare Books* which launches from July 2016, and approaches to diverse MOOC audiences, regardless of their baseline differences in ethnic, regional, or educational.i The MOOC program runs for three weeks and features the collection from the Institute of Oriental Classics as well as the visual materials from the Keio University Library collection. The course covers the various fields in bibliographical studies, such as bookbinding styles, types of manuscripts and illustrated books, and the history of book publishing in Japan. Along with these course topics, the aim of this research is to design and implement an online digital collection, which allows general audiences at different level to access and
interact with the vast collection from Keio University, using the combination of computational analysis and narrative visualization methods to provide a deeper understanding of pre-modern Japanese books.

The design process for developing aesthetically pleasing yet insightful digital collection is high dimensional and inherently complex. Methods and tools are widespread in the scholarly community, not only in the scientific disciplines but also in the humanities within the framework of digital humanities. However, the most important area in digital collection is the quality and efficacy of its design. Effective design and experience must be accessible to a plurality of people, and hence

this research advances the discussion with integrating digital curation strategies and narrative visualization format to the design in aiming to provide effective and intuitive experience for the diverse audience.

Through digitizing and visualizing temporal data in a narrative format, and focusing on both verbal and nonverbal aspects of the books, this research allows general audiences to interact with its diverse elements of Japanese culture, from micro to macro level. The implementation of digital collection provides practical and comprehensive insights of Japanese culture through books. Furthermore, this paper expects to prove that gaining new insights through historical temporal data does not only require technological advancement, but also an appropriate transformation and interpretation of the data through the combination of computational and visualization approach.

---

i FutureLearn, Japanese Culture Through Rare Books, https://www.futurelearn.com/courses/japanese-rare-books-culture (May 2016.)

[Invited Poster Presentation] Approach to Networked Open Social Scholarship

Ray Siemens (University of Victoria) and the INKE Research Group

As elements of our digital scholarly ecosystem continue to expand and evolve, there is an increasing necessity to serve both expert and public need for open access to information. The ubiquity of mobile technologies, the development of augmented reality, virtual reality, and location-based technologies, the challenge and influence of big data and, increasingly important, broad public participation in the production and use of digital knowledge repositories — these exemplify areas of challenge that present opportunities for those working in the area, toward leveraging these technologies and creating shared and integrated digital environments that will engage and benefit everyone, expert and general public alike. In this context, this poster presentation explores the next steps of the Implementing New Knowledge Environments Partnership for Networked Open Social Scholarship (INKE; inke.ca), itself united by the goal to explore, research, and build environments for open social scholarship in Canada and beyond, enhancing national and international research, digital infrastructure, and dispersed resources to develop innovative publishing and communication environments that connect those who share need for access to the information produced by our academic communities."

Verifying the Authorship of Saikaku Ihara's Kousyoku Gonin Onna

Ayaka Uesaka (Organization for Research Initiatives, Doshisha University)

INTRODUCTION

Saikaku Ihara (c.1642～93) was a haikai poet and fiction writer of the Genroku period (1688～1704) in Japan. After publishing the maiden works of *Koushoku ichidai otoko* (Life of a Sensuous Man;1682), he became the leading author of Ukiyozoushi. In the late eighteenth century, there was a Saikaku revival, inspiring many modern Japanese writers. Saikaku's works are known for their significance in developing Japanese novels today (Emoto and Taniwaki, 1996).

In this paper, we focus on *Kousyoku gonin onna* (Five Sensuous Women;1686). This work is well known from Saikaku's work. According to Teruoka (1949), *Kousyoku gonin onna* did not have a preface, signature and epilogue but it must be Saikaku's work. Tsutsumi (1957) mentioned that *Kousyoku gonin onna* did not have a signature but it was evident it was Saikaku's work. Emoto (1984) also has argued that *Kousyoku gonin onna* did not have a preface, signature and epilogue but it is recognized as Saikaku's work, and I agree with the opinion. These researchers stated *Kousyoku gonin onna* was written by Saikaku but there is no evidence to support it is Saikaku's work. The first edition of the work did not have a preface, epilogue, handwritten signature and signature seal, namely it is not described that *Kousyoku gonin onna* was written by Saikaku. Moreover, Kigoshi (1996) stated that particular information should be stated about the uncertainty of author because material did not exist that described *Kousyoku gonin onna* was Saikaku's work before Meiji period (1868～1912).

The aim of this paper is to evaluate the writing style of *Kousyoku gonin onna* using quantitative analysis. In this paper, we investigate Saikaku's twenty-four novels. A comparison was needed in order to more accurately characterize the Saikaku's writing style. In this research we also used Saikaku's student Dansui Houjyou (1663～1711) 's three novels.

DATASET

Saikaku's database was developed with his researchers, who are editors of *Shinpen Saikaku Zenshu* (Shinpen Saikaku Zenshu Henshu Inkai, 2000). Since Japanese sentences are not separated by spaces, we added spaces between the words in all of the sentences. In addition, information was added for the analysis. We also used Dansui's database for comparison, which is developed by Professor Hidekazu Banno and Professor Takayuki Mizutani. TABLE 1 shows a list of works in our database and the number of words in each work. According to our database, there are 572,231 words contained in twenty-four of Saikaku's works and 53,172 words contained in three of Dansui's works.

Table 1. Work name and the number of words

Saikaku's works		
Kousyoku ichidai otoko 36,781 words	*Shoen Okagami* 45,753 words	*Wankyu Isse no monogatari* 7,702 words
Kousyoku gonin onna 20,184 words	*Kousyoku ichidai onna* 26,581 words	*Saikaku shokoku hanashi* 16,444 words
Honchou nijyu hukou 18,419 words	*Nanshoku ookagami* 50,452 words	*Budou denraiki* 49,019 words
Kousyoku seisui ki 20,866 words	*Hutokoro suzuri* 22,839 words	*Nihon eitaigura* 29,547 words
Irozato motokoro setai 11,895 words	*Buke giri monogatari* 21,456 words	*Arashi ha mujyou monogatari* 8,727 words
Shin kashou ki 25,157 words	*Honchou nijyuu hukou* 26,466 words	*Seken hume zanyou* 21,260 *words*
Ukiyo eiga ichidai otoko 22,576 words	*Saikaku oki miyage* 17,204 words	*Saikaku oridome* 29,617 words
Saikaku zoku turezure 13,966 words	*Yorozu no humihougu* 16,940 words	*Saikaku nagori no tomo* 12,380 words
Dansui's works		
Shikidou otsuzumi 11,494 words	*Chuya youjin ki* 21,508 words	*Budou hariai okagami* 20,170 words

ANALYSIS AND RESULT

We examined the appearance rate of the particles and auxiliary verbs. These variables have a high appearance frequency and do not relate to the contents of a work.

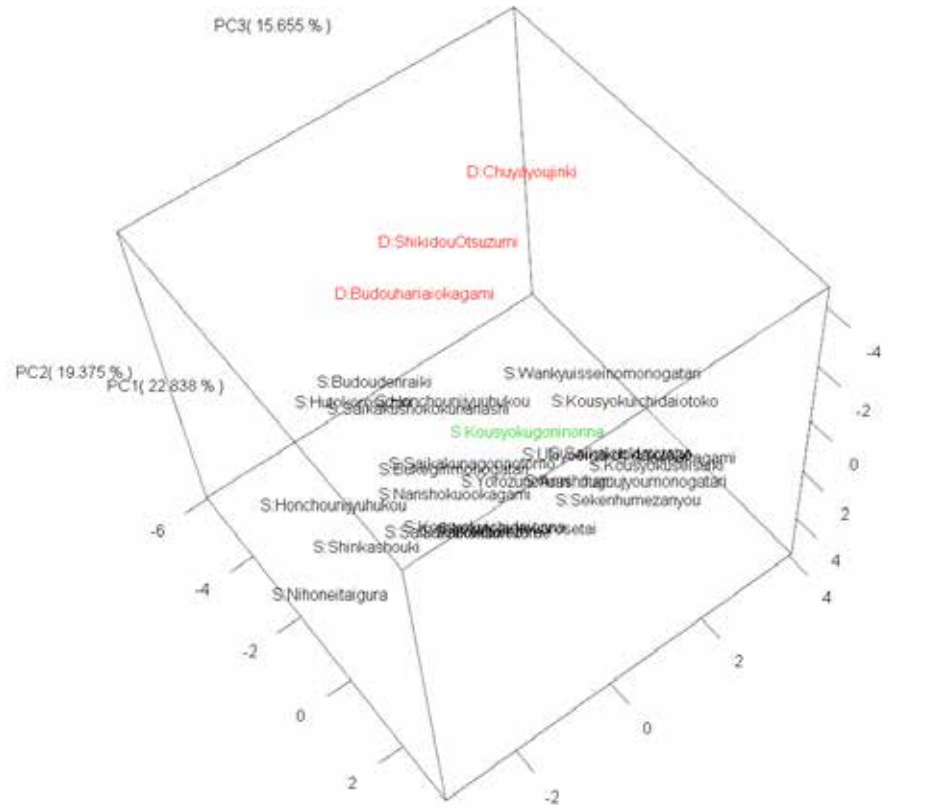

Figure 1. PCA results of the nineteen particles (95.575% of all the particles) for Saikaku's works and Dansui's works

FIGURE 1 shows the results of the analysis on the appearance rate of the particles using the Principal Component Analysis (PCA). PCA reduces the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much of the variation present in the

data set as possible (Jolliffe, 2002). When applied to the frequencies of high-frequency items in texts, PCA often successfully reveals the authorial structure in a data set (Kestemont et al., 2013). The proportion of variance of the first principal component is 0.22838, it is 0.19375 for the second, while it is 0.15655 for the third; the cumulative proportion up to the third principal component is 0.57868. In this figure, *Kousyoku gonin onna* is in close proximity to the other Saikaku's works. This result revealed that Saikaku and Dansui's works differed in the appearance rate of the particles. *Kousyoku gonin onna* was far from Dansui's works. Furthermore, we obtained similar result of the auxiliary verbs.

CONCLUSION

We conducted the quantitative analysis among Saikaku twenty-four works and Dansui three works. This result revealed that *Kousyoku gonin onna* possessed the same characteristics as Saikaku's works. From that viewpoint, *Kousyoku gonin onna's* author is Saikaku. In this study, we used Saikaku's works and Dansui's works as datasets and particles and auxiliary verbs as variables. Thus, we need to analyze and compare this issue to the other author's works and variables.

ACKNOWLEDGEMENTS

We would like to thank Professor Masakatsu Murakami, Professor Hidekazu Banno and Professor Takayuki Mizutani for their help on our research.

REFERENCES

[1] Emoto, Y. and Taniwaki, M. (1996). Saikaku Jiten. Ouhu.
[2] Teruoka, Y. (1949). Teihon Saikaku Zenshu Vol.2. Explanation Kousyoku gonin onna. Chuo Koron Shuppan.
[3] Tsutsumi, S. (1957). Nihon Koten Bungaku Taikei Saikaku Jyo. Explanation Kousyoku gonin onna. Iwanami Shoten.
[4] Emoto, H. (1984). Explanation Kousyoku gonin onna. Koudansha Gakujutsu Bunko.
[5] Kigoshi, O. (1996). The Uncertainty of the Authorship: Who Should Decide Koshoku-gonin-onna Belong to Saikaku?. Nihon Bungaku Vol. 45 No.10. pp.59～69.
[6] Shinpen Saikaku Zenshu Henshu Inkai. (2000). Shinpen Saikaku Zenshu. Bensei shuppan.
[7] Jolliffe, I.T. (2002). Principal Component Analysis. New York: Springer.
[8] Kestemont, M., Moens, S., and Deploige, J. (2013). Collaborative authorship in the twelfth century: A stylometric study of Hildegard of Bingen and Guibert of Gembloux. Literary and Linguistic Computing. pp.1～26.

Quantitative Analysis for Division of Viola Parts of Mozart's symphonies

Michiru Hirano (Tokyo Institute of Technology)

Introduction

This study focuses on the fact that some of the symphonies by Wolfgang Amadeus Mozart (17561791) include two viola parts. More specifically, the goal of the study is to examine whether separating out viola parts influences the orchestration of violin parts.
Symphony refers to a genre of orchestral composition, that has been actively composed since the eighteenth century [1]. Mozart composed in excess of forty symphonies throughout his life [2]. While symphonies are usually composed for orchestras comprised primarily from four string parts (two violin parts, a viola part and a cello part), some of Mozart's symphonies require two viola parts. We refer to this phenomenon as a separation of the viola parts within this paper. While acknowledging that the notion of separating the violas within symphonies is not common, still, there has been little discussion of its potential significance and of what Mozart might have been pursuing.
Even for works that contain two viola parts, the parts are frequently played the same and only rarely are the individual notes assigned to the respective viola parts. On the other hand, violins that are basically assumed to be played two distinct parts, are occasionally played together. Separating the violas means that the number of parts increases. For sections where violas are separated, if the ratio for the separation of violins is higher than usual, then separating the violas seem to intentionally increase the number of parts. In contrast, if the ratio is lower, then separating the violas does not imply an intention to increase the number of parts, but rather to give the violas the roles that violins ought to have. If the ratio does not change when violas are separated, then, separating the violas does not influence the orchestration of the violins, which would imply another objective. This study utilizes computational methods to examine whether separating the violas influences the ratio of separation for violins.

Method

There are 17 Mozart symphonies where the initial movement is divided into two viola parts, and this study targets those 17 initial movements.
The following procedure is done for each of the 17 works. First, the scores were obtained from "The New Mozart Edition" by B¨arenreiter Vertrag, which contains the most authoritative scores for all of Mozart's compositions currently available[1]. Next, the scores were exported into the MusicXML format, which is a textural representation of the musical notation suitable for digitization. Then, every measure is examined to determine whether or not paired parts (for both violins and violas, respectively) are consistent. Consistency for paired parts means that they are not separated, which inconsistency means that the parts are separated. The durations and pitches of notes are used in determining consistency. If any differences in terms of note durations or pitch are observed within a measure, the parts of the measure would be regarded as being inconsistent. Notes with pitch belonging to the same pitch class, however, are regarded as being consistent, even if there is a gap between octaves. Every measure was examined for the correspondences between consistencies and inconsistencies for the violins and the violas and the frequencies of measures falling under the various conditions are listed in Table 1. Finally, Fisher's exact tests were conducted to identify whether any significant differences exist between the ratios of A (both violins and violas are separated) to B (violins are not separated but violas are) and between the ratios C (separated violins but violas are not separated) to D (neither violins nor violas are separated) in Table 1.

[1] Andr´e Hodeir. Les formes de la musique. Presses Universitaires de France, 1951. ([In Japanese.] Ongaku no keishiki [The forms of the music], Hidekazu Yoshida, trans.,Hakusuisha(1973)).

Results

Table 2 presents the separation ratios and p-values obtained from the Fisher's exact tests. There are 11 of the 17 works that have p-values that are lower than the 0.05 significance level (K.43, 112, 114, 132, 173dB, 189k, 385, 425, 543, 550, and 551). Thus, for those works, it is possible to reject that null hypothesis that the separation ratio is not influenced by separating the violas parts.

Discussion

The analyses results failed to observe significant relations between the ratios for separating violins and violas for six of the 17 works (K.133, 162, 173dA, 319, 338, 504). Moreover, of the 11 works for which significant differences between the ratio for separating violins when violas are separated, five works (K.43, 173dB, 385, 425, 543) have greater ratios between A and B compared to the ratio between C and D. For those works, separating the violas would seem to intentionally increase the number of parts. For the remaining six works (K.112, 114, 132, 189k, 550, 551), however, there are higher ratios between B and A. In those cases, separating the violas would appear to inhibit any separation for violins. Accordingly, it would seem that Mozart did not have a single reason for separating the viola parts, and the objective varied across different works.

Conclusion

This study conducted a quantitative analysis of Mozart's symphonies that include two viola parts. Specifically, we examined whether the ratio for separation of the violin parts is influenced when the viola parts are separated. Such influences were found to be significant for 11 of the 17 works. Five works exhibited a tendency for the violins to be separated more frequently when the violas are separated, with the opposite trend observed in the remaining works. Consequently, it would seem that Mozart had different objectives in mind when he separated the viola parts of his symphonies.

References

[1] Andr´e Hodeir. Les formes de la musique. Presses Universitaires de France, 1951. ([In Japanese.] Ongaku no keishiki [The forms of the music], Hidekazu Yoshida, trans.,Hakusuisha(1973)).

Table 1: Frequency distribution for the correspondences between the consistencies and inconsistencies for violins and violas: The cells labeled A, B, C and D indicate the numbers of measures conforming to the respective conditions.

	Violins are separated (inconsistent)	Violins are not separated (consistent)
Violas are separated (inconsistent)	A	B
Violas are not separated (consistent)	C	D

Table 2: List of materials and their measured values: The number assigned to each work is from the sixth edition of the K¨ochel catalogue, which is a chronological catalogue of Mozart's compositions. Items labeled A, B, C and D correspond to the conditions presented in Table 1. The rightmost item is the p-value derived from relevant Fisher's exact test.

Work	A	B	C	D	p-value
K.43	42	2	43	14	0.01
K.112	5	9	81	29	0.01
K.114	2	5	120	12	0.00
K.132	22	17	95	14	0.00
K.133	16	0	148	18	0.37
K.162	21	9	90	15	0.06
K.173dA	2	1	91	52	1.00
K.173dB	16	0	95	103	0.00
K.189k	4	10	124	35	0.00
K.319	64	18	226	62	1.00
K.338	16	13	128	107	1.00
K.385	11	0	126	67	0.02
K.425	26	1	180	80	0.00
K.504	6	3	221	72	0.69
K.543	17	0	79	213	0.00
K.550	1	48	140	110	0.00
K.551	0	4	194	115	0.02

[2] Neal Zaslaw. Mozart's symphonies: context, performance practice, reception. Oxford University Press, 1989. ([In Japanese.] Mozart no symphony: context, ensou jissen, juyou, Tadashi Isoyama and Miho Nagata, trans.,Tokyo shoseki(2003)).

Characteristics of a Japanese Typeface for Dyslexic Readers

Xinru Zhu (University of Tokyo)

Introduction

Evidence shows that 3%–5% of the population have developmental dyslexia[1] in Japan [1], and providing them with assistive environment is essential. While it is held that typefaces have impacts on dyslexic readers [2], Japanese typefaces for dyslexic readers have not been created, mainly because it is not easy to provide a special typeface that fits everyone with dyslexia.
Against this backdrop, we are developing (i) a Japanese typeface for people with developmental dyslexia and (ii) a typeface customization system, targeting the situation in which people read articles or textbooks.
This poster presents the Japanese typeface we designed for dyslexic readers. In designing the typeface, we analysed Latin typefaces designed for dyslexic readers and extracted characteristics they have, defined desiderata for Japanese typefaces for dyslexic readers by mapping these characteristics to Japanese char- acters, and created a Japanese typeface for dyslexic readers by applying these desiderata for dyslexic readers. We elaborate on each of these steps in our pre- sentation.

Characteristics of Latin Typefaces for Dyslexic Readers

There are several Latin typefaces specially designed for dyslexic people, includ- ing Dyslexie, OpenDyslexic, Lexie Readable, Sylexiad and Read Regular. We examined the characteristics of Dyslexie, OpenDyslexic and Lexie Readable for the reason that they are relatively widely used and evaluated in several studies. Studies show that typefaces have significant impacts on readers with dyslexia [3] and with specially designed typefaces, dyslexic readers either was able to read with less errors [4, 5, 6] or preferred the specially designed typefaces compared to normal typefaces [7].
In order to identify the characteristics of the special designed typefaces, we measured[2] the letterforms of 3 special typefaces and 6 normal sans-serif type- faces[3] and summarized them parametrically based on PANOSE classification[4], numerically based on the sizes and ratios of the typefaces and visually based on the direct comparison. The font data was converted to the Unified Font Object[5] from commonly used format to make it easy to access to coordinates of points constructing glyphs from Python scripts. The methods adopted ensure repro- ducibility and objectivity of the study.
Table 1 describes PANOSE numbers and the characteristics of typefaces they show. Table 2 and Table 3 show the PANOSE values of Arial and Dyslexie and Figure 1 and Figure 2 shows the average sizes and ratios of the typefaces. Figure 3 is a part of the visual comparison of Arial and Dyslexie in the same size, in which blue letters are in Arial and red ones are in Dyslexie.
The results show that Latin typefaces for dyslexic readers have the following characteristics.

1. The characteristics of the entire typeface:
   (a) Rounded sans-serif typefaces,
   (b) Larger letters in the same size,
   (c) Larger height/width ratios,

[1] Developmental dyslexia is defined as “a specific learning disability that is neurobiological in origin. It is characterized by difficulties with accurate and/or fluent word recognition and by poor spelling and decoding abilities” according to the International Dyslexia Association.
[2] Measurements and modification of typefaces were conducted using the programming lan- guage Python and RoboFont, a Python based font editor (http://doc.robofont.com/).
[3] They are Arial, Calibri, Verdana, Trebuchet, Comic Sans, and Sassoon Primary. These type- faces are selected based on the recommendation of the British Dyslexia Association.
[4] PANOSE is “a system for describing characteristics of Latin fonts that is based on calculable quantities” [8].
[5] The Unified Font Object is a human readable XML format for storing font data. http:// unifiedfontobject.org/.

(d) Standard x-heights,
(e) Longer descenders and ascenders,
(f) Bolder strokes,
(g) Contrast in stroke width.

2. The characteristics related to identifying similar letters:
(a) Similar letters slanted or rotated to opposite directions,
(b) Uppercase “I, J” and numeric character “1” with serifs,
(c) Numeric character “0” with a dot inside the counter,
(d) Asymmetry letterforms of lowercase “p, q” and “b, d”,
(e) Handwritten style of lowercase “a, y” and numeric character “9”,
(f) Larger counter sizes of lowercase “a, c, e, s”.

Table 1: PANOSE Number and Characteristics of Typefaces

PANOSE Number	PANOSE Values of Arial
1	2: Latin font for running text and titling
2	11: normal sans serif
3	5: book
4	4: even width
5	2: no contrast
6	2: no variation
7	3: straight arms / wedge termination
8	2: normal / contact
9	2: standard midlines / trimmed apexes
10	4: constant letters / large x-height

Table 2: PANOSE Values of Arial

PANOSE Number	Characteristics
1	Family Kind
2	Serif Style
3	Weight
4	Proportion
5	Contrast
6	Stroke Variation
7	Arm Style and Termination of Open Curves
8	Slant and Shape of the Letter
9	Midlines and Apexes
10	X-height and Behavior of Uppercase Letters Relative to Accents

Table 3: PANOSE Values of Dyslexie

PANOSE Number	PANOSE Values of Dyslexie
1	2: Latin font for running text and titling
2	11: normal sans serif
3	6: medium
4	3: modern
5	3: very low
6	5: gradual / vertical
7	7: nonstraight arms / horizontal termination
8	2: normal / contact
9	2: standard midlines / trimmed apexes
10	3: constant letters / standard x-height

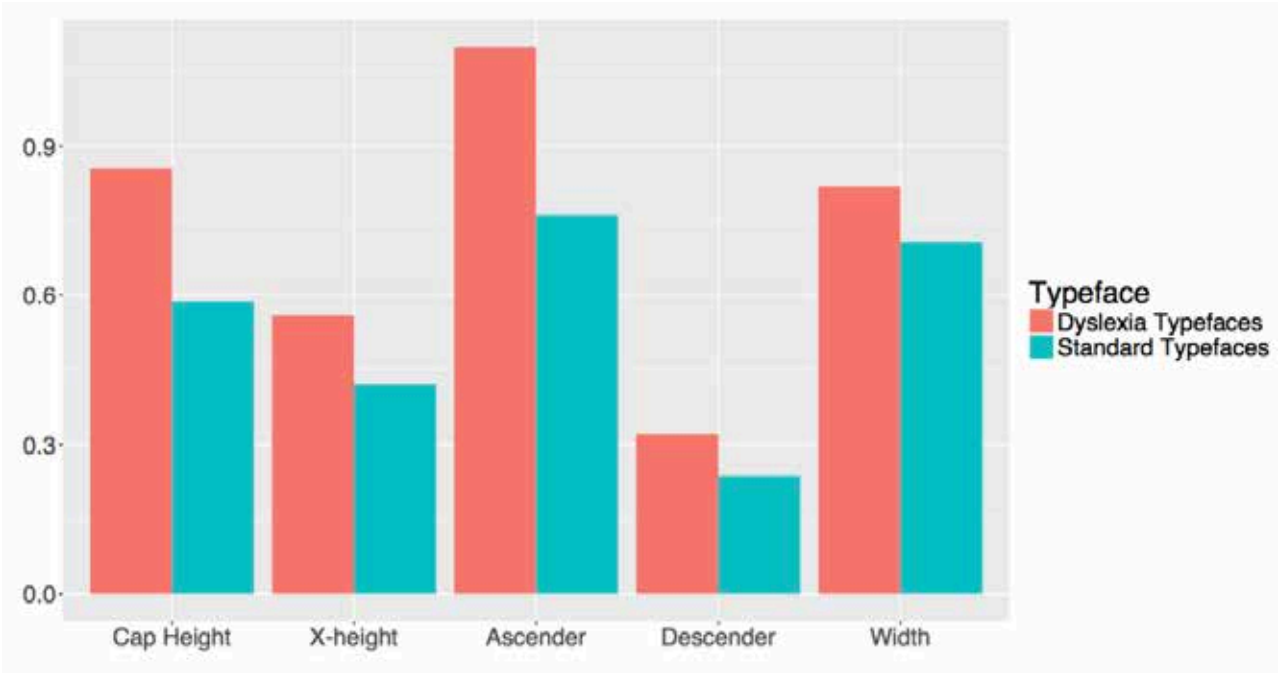

Figure 1: Sizes of the Typefaces

Desiderata for Japanese Typefaces for Dyslexic Readers

A Japanese font set includes Latin characters, Kana characters and Kanji char- acters, not mentioning punctuation marks and other symbols, in which Latin characters and Kana characters are phonograms while Kanji characters are lo- gograms [9]. Neuropsychological studies indicate that phonograms and logograms are processed differently in human brains [10], which makes it reasonable to dis- cuss possible characteristics of Kana characters and Kanji characters separately.

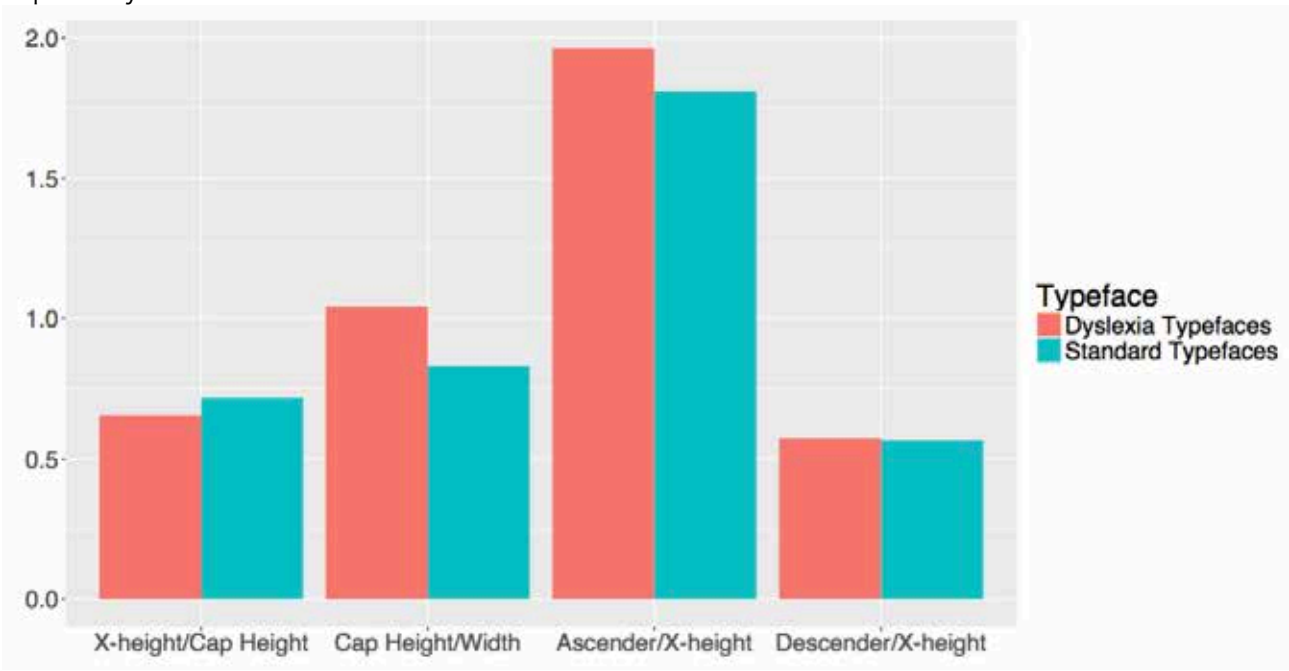

Figure 2: Ratios of the Typefaces

Since Kana characters are phonograms same as Latin characters, the hypoth- esis is that some characteristics of the Latin typefaces for dyslexic readers can be applied directly to the entire Kana typeface. It is indicated that forms of some Kana characters are similar to one another which leads to confusion during char- acter recognition [11]. The characteristics related to identifying similar letters hence can be applied to those characters. The possible characteristics of Kana typefaces are listed below.

1. The characteristics of the entire typeface:
   (a) Maru gothic typefaces[*6],
   (b) Larger characters in the same size,
   (c) Larger height/width ratios,
   (d) Bolder strokes,
   (e) Contrast in stroke width,
   (f) Larger counters.
2. The characteristics related to identifying similar characters:
   (a) Hiragana characters “ら，う”, “る，ろ”, “は，ほ” [11], “い，こ”, “め，ぬ”, and “へ，く” [12] modified distinguishable,
   (b) Katakana characters “ス，ヌ”, “セ，ヤ”, “ウ，ワ”, “ワ，フ”, “ワ，サ”, “ソ，ン” and “ユ，エ” [11] modified distinguishable.

As for Kanji characters, there are two possible strategies. First, Kanji charac- ters can be treated in the similar way as Kana characters since the visual aspects of Kanji characters are considered to play an important role in character recog- nition [13]. The second strategy is to emphasize the structure of Kanji characters inside the typeface according to widely adopted assistive practices.

Figure 3: Visual Comparison of Arial and Dyslexie

A Prototype of Japanese Typefaces for Dyslexic Readers

We selected all the Hiragana and Katakana characters and 80 Kanji characters instructed to be taught in the first grade in elementary schools by the Ministry of Education, Culture, Sports, Science and Technology of Japan to be included in the first prototype of the typeface. Since each Kanji character is constructed with certain strokes, the idea is to start from the characters with fewer strokes and expand gradually. Kanji characters will be expanded to 2136 characters of Jo¯yo¯ Kanji, commonly used Kanji characters announced by the Government of Japan, in the final design.

The first prototype of the Japanese typefaces for dyslexic readers is modi- fied based on an open source Japanese typeface. We converted it to the Unified Font Object and applied the possible characteristics summarized above by run- ning Python scripts on the data of glyphs. The results of modification will be demonstrated in the poster. The prototype will be put on evaluation in cooperation with dyslexic readers in further studies and the results will be reflected to the characteristics of the Japanese typefaces for dyslexic readers.

References

[1] Tomonori Karita, Satoshi Sakai, Rumi Hirabayashi, and Kenryu Nakamura. Trends in Japanese Developmental Dyslexia Research [in Japanese]. Journal of Developmental Disorder of Speech, Language and Hearing, 8:31–45, 2010.

[2] Shinji Iizuka. A Classification of Assistive Technologies for Reading Disor- der Based on the Process of Language Understanding [in Japanese]. IEICE Technical Report. Welfare Information Technology, 106(612):43–48, 2007.

[3] Luz Rello and Ricardo Baeza-Yates. Good Fonts for Dyslexia. In Proceed- ings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility, page 14. ACM, 2013.

[4] Maya Grigorovich-Barsky. The Effects of Fonts on Reading Performance for Those with Dyslexia: A Quasi-Experimental Study, 2013.

[5] Tineke Pijpker. Reading Performance of Dyslexics with a Special Font and a Col- ored Background. Master thesis, University of Twente, 2013.

[6] Renske De Leeuw. Special Font For Dyslexia? Master thesis, University of Twente, 2010.

[7] Robert Alan Hillier. A Typeface for the Adult Dyslexic Reader. PhD thesis, Anglia Ruskin University, 2006.

[8] Yannis Haralambous. Fonts & Encodings. O'Reilly Media, 2007.

[9] Florian Coulmas. Writing Systems: An Introduction to Their Linguistic Analy- sis. Cambridge University Press, 2003.

[10] Makoto Iwata and Mitsuru Kawamura. Neurogrammatology [in Japanese]. Igaku-Shoin, 2007.

[11] Tatsuya Matsubara and Yoshiro Kobayashi. A Study on Legibility of Kana- letters [in Japanese]. The Japanese Journal of Psychology, 37(6):359–363, 1967.

[12] Nobuko Ikeda. Research on Educational Support of Japanese Language Learners with Developmental Dyslexia [in Japanese]. Journal of the Study of Japanese Language Education Practice, (2):1–15, 2015.

[13] Cecilia W. P. Li-Tsang, Agnes S. K. Wong, Linda F. L. Tse, Hebe Y. H. Lam, Viola H. L. Pang, Cathy Y. F. Kwok, and Maggie W. S. Lin. The Effect of a Visual Memory Training Program on Chinese Handwriting Performance of Primary School Students with Dyslexia in Hong Kong. Open Journal of Therapy and Rehabilitation, (3):146–158, 2015.

Digitally Archiving Okinawan Kaida Characters

Mark Rosa (Ph. D., University of Tokyo, 2016)

The native Okinawan kaida writing system, created in the Yaeyama islands in the 17th to 19th centuries to track tax payments and record family holdings and contributions, and developed most highly on Yonaguni at the end of this period, has never been encoded digitally. This short paper will use two newly-discovered records, one stored in the archives of the National Museum of Ethnology in Suita, Osaka, and another in the library at the University of the Ryukyus, Okinawa, as a sample of the kinds of texts that digital encoding can be valuable for.

"Full writing," in which any verbal utterance can be expressed, was never developed for the various languages of the Okinawan islands. A system of partial writing called sūchūma was used for simple tallies of money, food, firewood, and other items, and combined with families creating symbols (called yaban on most islands and dahan on Yonaguni) to indicate their names, made basic record-keeping possible. In the southwesternmost islands – the Yaeyamas and Yonaguni – glyphs were devised for animals and foodstuffs, creating the kaida writing system in which more detailed records became possible: names, dates, items taken or possessed, and numbers.

The number of available samples of kaida writing is still – and might always be – small. The system began to fall out of favor when the first Japanese school was built on the island in 1885, and declined further when the hated capitation tax came to an end in 1903. The last reports of active use of this system date from the 1920s, and today only a small handful of islanders, all born around this time or earlier, can remember how to write it even partially: one such is Nae Ikema, born in 1919 and aged 96 at the time of writing. (Many more islanders of all ages can write their families' dahan.)

No attempt has previously been made to encode these characters so that they can be preserved and transmitted digitally. The more primitive sūchūma, being basic shapes such as circles, squares, triangles, crosses, and lines, could conceivably be covered by existing Unicode characters, but the numerals are distinctive enough from Japanese/Chinese to warrant their own encoding, and the pictographs are unlike anything seen in those two languages.

This work will introduce a TrueType font for kaida characters, created by the author, and will explore the above-mentioned records from the University of the Ryukyus and National Museum of Ethnology and attempt to recreate them digitally. The addition of private individuals' dahan in the Private Use Area will be necessary for the records to be complete.

The next stage will be to make ordinary speech digitizable by creating an input method editor for the language in general, written in today's Japanese- based kanji and kana, and not just the historical writing system. This presentation will conclude with a brief introduction to this future step.

Keywords
kaida writing, yonaguni, native okinawan writing, partial writing, unicode

Attributes of Agent Dictionary for Speaker Identification in Story Texts

Hajime Murai (Tokyo Institute of Technology)

Introduction

In order to interpret and to analyze story structure automatically, it is necessary to identify who the agents are that appear in the story. This involves identifying general expressions in story text for story agents and analyzing pronouns, omissions, and the aliases of agents.
These goals assumes use of natural language processing techniques such as morphological analysis [1] and dependent analysis [2]. After morphological information and dependent relationships were obtained, the next step would be identification of agents and those behaviors in order to analyze the narratological structure of the story texts.
In this article, agents in story texts are generally proactive beings who have a will, though there may be some exceptions. In many cases, the agents are human beings. However, there are also various other agents, such as aliens, space creatures, devils, ghosts, robots, and automated machines, depending on the genre of the stories.
In general texts, some agents may be called by proper nouns at first time. However in many cases, they would be called by pronouns after second time. Moreover, most of agents have several aliases as a nickname, an official position, or a role in the family. Therefore it is necessary to identify the relationships between proper nouns and pronouns and other expressions about agents in a story text.
Moreover in Japanese text, the omission of agent vocabulary in sentences occurs frequently. Therefore, it is also necessary to estimate the omitted agent words in order to extract the story structure. In addition to that, the speaker and listener are not clarified in the dialogue texts of many stories. In such cases, the estimation of agents is also necessary.

Attributes for Agent Estimation

These estimation tasks regarding agents are very complex and the accuracy of the results is not sufficient even with recent technologies [3]. However, there are some clues to identify those agents. At first, types of pronouns give information about referring agent words. For example, “He” signifies that referred agent is male and singular. For instances, if there is “He” in some text and also if there is only one male singular proper noun, that “He” probably matches to the male singular proper noun.
In addition to those, honorific expressions are frequently appeared in dialogues in story texts. If hierarchical relationships between appeared agents in some story text can be extracted, honorific expressions become important clue to estimate and to identify agents. Moreover, calling expression such as “Honey” in dialogue also show relationships between agents. Therefore, general knowledge about relationships between agents should be stored as some database for precise agent estimation.
For instance, there are agent words in story texts that indicate family relationships (father, mother, sister, brother, etc.), vocational relationships (president, employee, etc.), and general nature of relationships (enemy, ally, friend, etc.). In some stories, it is not only individuals but also specific groups, organizations, regions, states, tribes, and nations that become agents. At first those agent words should be collected and should be categorized. In the next step, attributes for agent estimation could be granted to those words.
Table 1 shows current list of necessary attributes for agent estimation. It is desirable to extract those attributes from some elements in story texts.

Table 1: Attributes and Potential Clues for Agent Estimation

Attributes Types	Potential Clues
Gender	Proper noun, ponoun, suffix, particle
Singular / plural	Proper noun, ponoun, suffix
Polite / rude	Verb, prefix, ponoun, suffix, particle
Family relationship	Agent word noun
Workplace relationships	Agent word noun
Social status	Agent word noun

Structures for Agent Dictionary

In order to utilize attributes of agent words in agent estimation tasks, it is neccesary to construct some dictionary or database which contains those information about agent attributes.
As shown above, there is a wide range of agent vocabulary indicating proactive beings in the story text. Nevertheless, it is possible to extract these agent words from the story text and to construct a database list. Moreover, it may be possible to make a machine-readable, structured database based on the categorization of type of vocabulary and relationship.

Table 2: Category for Agent Words

Large Category	Small Category	Examples
Being	Person	Person, human, alien
	Biological	Creature, monster, behemoth
	Artificial	Robot, U.F.O., automatic machine
	Supernatural	Ghost, devil, demon, genie, Satan, Buddha
Relation	Biological	Brother, sister, son, daughter, parent, wife, husband, women, newborn, patient, handsome, beauty
	Profession	Doctor, attorney, employees, president
	Group	Senior, colleague, junior, representative
	Spatio-temporal	Ancients, Westerner
	General	Hero, enemy, friend, reader
Proper noun	Person	Sato, Suzuki, Ichiro
	Tribe	Earthlings, Indian, Asian
Pronoun	First	I, we
	Second	You, Thou
	Third	He, She, It
	Interrogative	Who

Table 3: Example of Attributes of Agent Words

Attributes Types		"Brother"	"Wife"	"Elderly"	"President"
Gender		Male	Female	Both	
Singular / plural		Singular	Singular	Singular	Singular
Polite / rude				Polite	Polite
Age				Old	
Social Status					High
Family	Generation	Same	Same		
	Distance	Near	Near		
	Biological / in-law	Biological	In-law		
	Companionship		Marriage		

Therefore, agent vocabulary appearing in story texts and general vocabulary from dictionaries that can be used as agent vocabulary were collected. The vocabulary was then categorized and a structured list of agent vocabulary was developed [4] (Table 2).
In addition to the category, attributes are granted to those collected agent words. Table 3 shows an example of stored attributes for each agent word. In table 3, agent words about family were granted attributes about family.

Conclusions and Future Works

In order to estimate relationships between agent words in story texts, relevant attributes were examined and those were structured with the category of agent words. By utilizing the developed database of agent vocabulary, candidates for text expressions which may indicate agents in story text can then be easily identified. If likely candidates for agents can be detected, they will become the foundation for more precise story structure analysis.

References

[1] Matsumoto Y, Kitauchi A, Yamashita T, Hirano Y, Matsuda H, Takaoka K, Asahara M. Japanese morphological analysis system ChaSen version 2.0 manual. NAIST Techinical Report. Apr. 1999.
[2] Daisuke Kawahara, Sadao Kurohashi. A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis, In Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, pages 176-183, June 2006.
[3] Hua He, Denilson Barbosa, and Grzegorz Kondrak. Identification of speakers in novels. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1312–1320, Sofia, Bulgaria, August 2013.
[4] Hajime Murai. Creating a subject vocabulary dictionary for story structure extraction. IPSJ Symposium Series, 2015:111–116, December 2015 (In Japanese).

Trends in Centuries of Words: Progress on the HathiTrust+Bookworm Project

Peter Organisciak, J. Stephen Downie (University of Illinois at Urbana-Champaign)

The HathiTrust+Bookworm (HT+BW) project is providing quantitative access to the millions of works in the HathiTrust Digital Library. Through a tool called Bookworm, digital humanities scholars can use outofthebox exploratory visualization tools to compare trends in all or parts of the collection, or use the API directly to query for more advanced questions. In this poster, we present the progress of the HT+BW project and discuss both its potential value to the digital humanities scholars and its current limitations.

HT+BW[1] is a quantitative text analytics tool built on top of the HathiTrust collection through improvements to a tool called Bookworm. HathiTrust, a consortium of library and cultural heritage institutions around the world, holds nearly 15 million scanned volumes, about 39% of which are in the US public domain. The current stage of HT+BW allows access to these public domain works, with ongoing work toward representing incopyright works and those of unknown status.

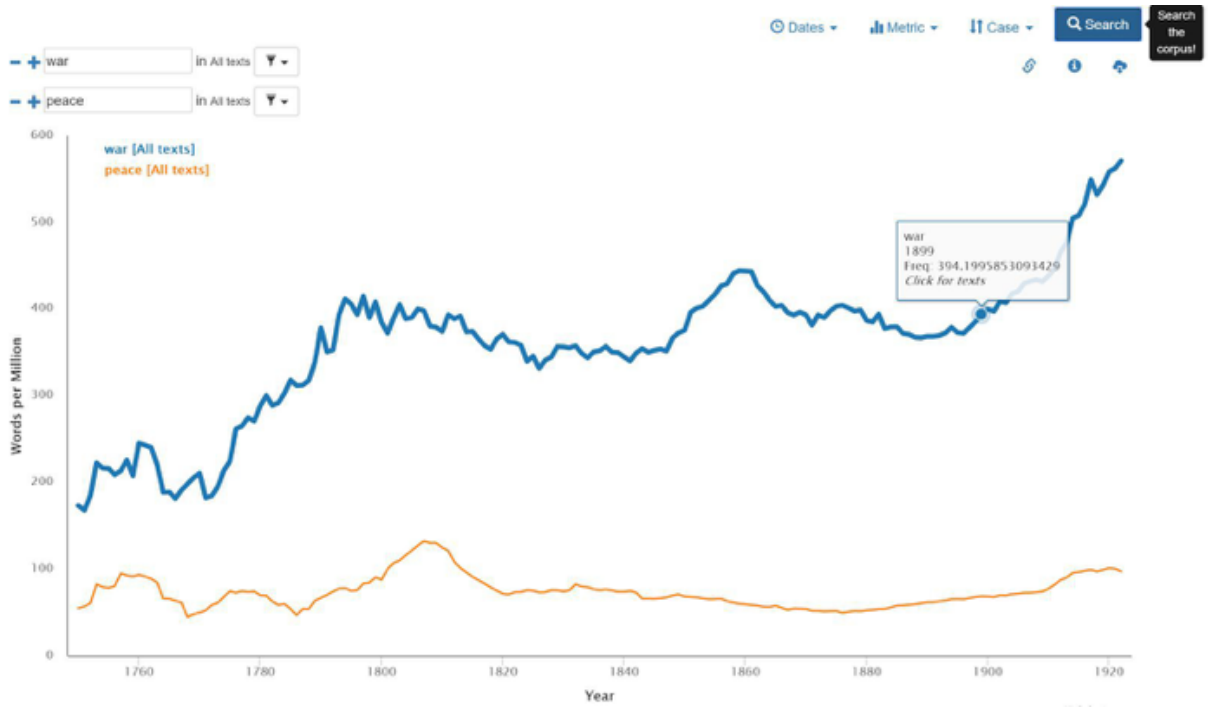

Figure 1: HT+BW in its simplest form: comparing different words over-time, corpus-wide

The tool underlying HT+BW is called Bookworm, a spiritual successor to the Google Ngrams Viewer (Michel et al. 2011). As with the earlier tool, the primary unit of analysis in Bookworm is the word token and the most common interface is a time series line chart.

Likewise, against the HathiTrust collection, the trends visualized also span centuries and millions of published works. However, HT+BW is significantly more robust than its popular predecessor: allowing more nuanced forms of inquiry, different visual interfaces for exploring results, and an application programming interface (API) that enables direct access to counts.

First, HT+BW can be queried by subsets of the data, rather than simply by year. Rather than only searching for trends of a word over time, one can compare that words trends for different classes of books, different genres, and different geographic provenance.

Faceting by metadata opens the door to much more nuanced questions. With HT+BW, one does not even have to use a word as a query: one could simply compare text counts between facets.

[1] [*1] http://bookworm.htrc.illinois.edu

For example: what subject areas are seen in texts published in the United States? What genres are popular in Japanese texts? How did the popularity of serials grow between countries?

Figure 2: Clicking on the visualization calls up links to the original works in the HathiTrust Digital Library

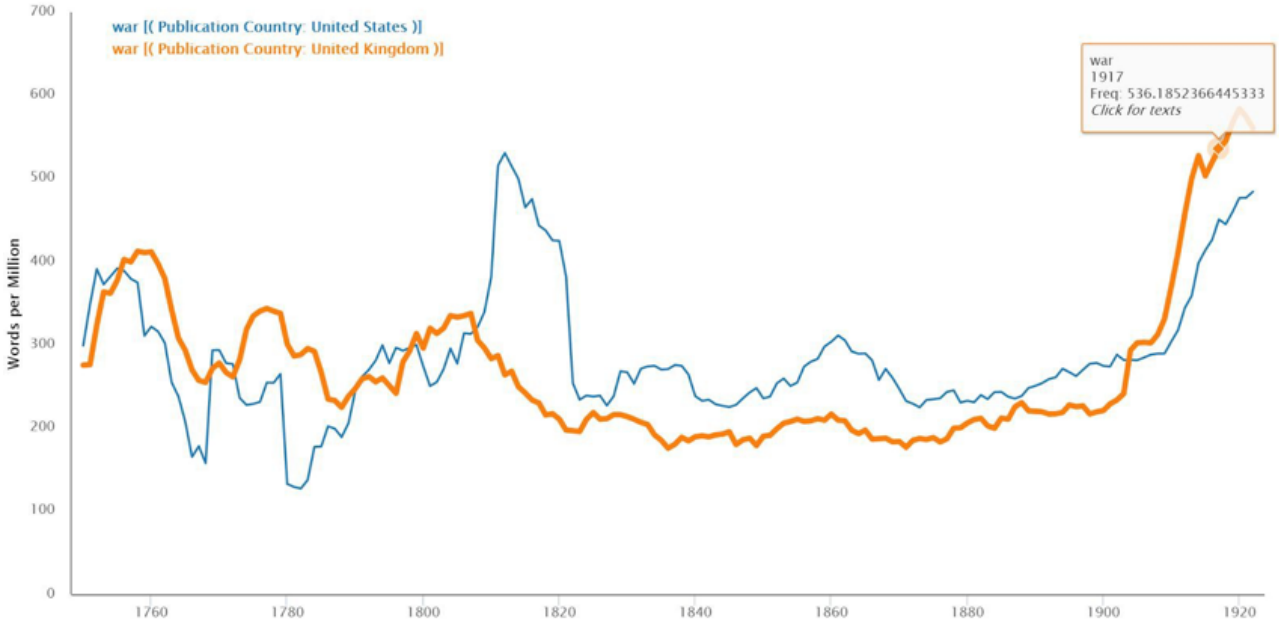

Figure 3: Comparing the same word over different subsets: it this case, books published in the US version versus those in the UK.

Another area where HT+BW moves beyond its antecedent is that not all questions need to be structured along years. Subsequently, visualization does not need to be structured as a time series line chart, and alternate visualizations are in development (Schmidt 2016).
However, the raw quantitative counts for highly customized queries can be returned using a public API, providing a path for scholars to move from exploration to more indepth questions.
HT+BW includes books from all around the world in 345 different languages. The materials held by HathiTrust are contributed to mainly from western institutions, meaning that English is the bestrepresented language in the collection, followed by other European languages. The best-represented Asian language is Japanese, with 73 thousand books, followed by Chinese with 32 thousand books. Bookworm supports extended Unicode characters, so Japanese is supported in

the various uses of HT+BW. One limiting factor for scholars working with Japaneselanguage texts is that their metadata and coverage will not be as strong as for betterrepresented languages. For example, nearly no Japanese texts in the current HT+BW have a subject class assigned.

The current coverage of the HT+BW is of public domain works, biasing the collection toward older works. This is a temporary limitation, and the ongoing project is prioritizing an expansion of the data to all 15 million works. Another limitation being addressed in future work is that current searches can only be done on single word phrases.

HT+BW provide quantitative, flexible access to the millions of texts in the HathiTrust Digital Library. Currently it supports single word queries against 4 million public domain works, with support for facets over a variety of metadata fields and even visualization of personal collections of texts. This poster describes the current state of the HT+BW, and outlines its future work in supporting more words for more books.

References

[1] Michel, JeanBaptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, et al. 2011. “Quantitative Analysis of Culture Using Millions of Digitized Books.” Science 331 (6014): 176–82. doi:10.1126/science.1199644.

[2] Schmidt, Benjamin M. 2016. "BookwormD3". Tool. Github. https://github.com/bmschmidt/BookwormD3.

Development of the Dictionary of Poetic Japanese Description

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

Introduction

The main purpose of this project is to de- velop a dictionary for Yamato Japanese description(Yamamoto et al. 2014). To this purpose, the present study proposes a method of extracting sub communities as classical Japanese poetic vocabu- lary. The analysis is based on co-occurrence pat- terns defined as any two words appearing in the same poem.

Many scholars of classical Japanese poetry have tried to explain constructions of poetic vocabulary based on their intuition and experience. As scholars can only describe constructions that they can consciously point out, those that they are un- conscious of will never be uncovered. When we de- velop a dictionary of poetic vocabulary using only our intuitive knowledge, the description will lack important lexical constructions. We believe that in order to conduct more exact and unbiased de- scriptions, it is necessary to use computer-assisted descriptions of poetic word constructions using co-occurrence weighting methods on corpora of classi- cal Japanese poetry. A typical item in a general dictionary con- tains the item's definition, part of speech, expla- nation, and example sentences. An item in the proposed dictionary contains not only the above-mentioned four types of information, but also in-cludes lists of words grouping sub communities, which allows one to better grasp the construction of poetic words.

In terms of lexical study, many quantitative studies of vocabulary are focused on the frequency of the occurrences of words. However, research re- lying on word frequency alone does not contribute to the analysis of mid-range words—words with not too high but not too low frequencies (Hodoˇsˇcek and Yamamoto 2013). We therefore use the R package 'linkcomm' to calculate network centrality between collocations (Freeman 1978). In the context of lexi- cal analysis, we regard this calculation of sub com- munity discovery as a way to describe the poetic roles of mid-range words.

Methods

We will attempt to extract all of the sub com- munities of ume (plum), sakura (cherry), and tachibana (mandarin orange) from the Hachidaishu¯ database[1]. We will use 'linkcomm' procedure to calculate word centrality to uncover the key sub communities (Csardi and Nepusz 2006, Ahn et al. 2010). As materials of this research we will use the Hachidaishu¯ (ca. 905–1205). We mainly collect the data from Kokkataikan (Shin-pen Kokkataikan Henshu¯ Committee 1996), Niju¯ichidaishu¯ database published by NIJIL (Nakamura et al. 1999), Shin- Nihon Koten Bungaku Taikei (Kojima and Arai 1989), and Shin-kokinshu¯ (Kubota 1979).

Results

Table 1 and Figure 1 were extracted based on the network of tachibana (mandarin orange). We found that the three methods, average, McQuitty, and single, are not different in terms of community discovery. We discovered the largest community, mukashi, (old times) which includes 15 nodes in the graph of tachibana.

Discussion

Table 1 lists the centrality values given by the three methods, which show similar tendencies among the three methods. These words are clearly relating to the poem which is famous for its

[1] We will report only on tachibana because of limited space.

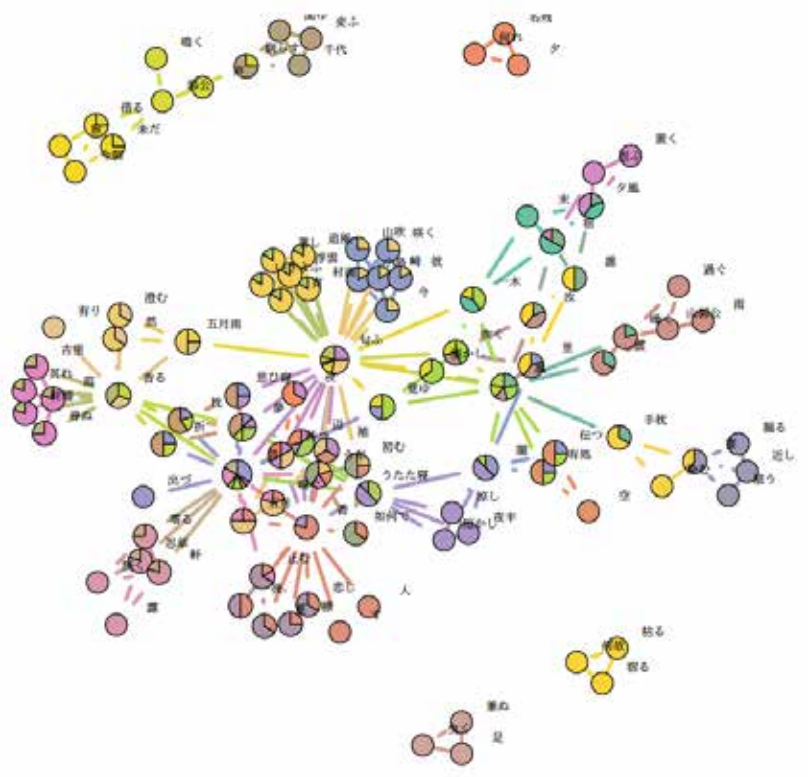

Figure 1: Network of tachibana (mandarin orange)

Table 1: The sub-cluster of tachibana (mandarin or- ange): Top 10 words having higher den- sity values are extracted; we used the aver- age, McQuitty, and single clustering meth- ods; values in parentheses indicate maxi- mum partition density.

	average (.43)		mcquitty (.43)		single (.38)	
No.	node	edge	node	edge	node	edge
1	mukashi	7	mukashi	7	mukashi	5
2	nihofu	6	nihofu	6	nihofu	4
3	kaze	5	kotoshi	4	yume	4
4	yume	5	atari	4	kaoru	3
5	kotoshi	4	matsu	4	kotoshi	3
6	atari	4	kaze	4	somu	3
7	matsu	4	yume	4	samidare	3
8	kaoru	3	somu	3	ori	3
9	samidare	3	kaori	3	makura	3
10	somu	3	yami	3	omohine	3

tachibana flowers[2] written by an anonymous author but commonly at- tributed to Ariwara no Narihira.

All poems have some supporting words sup- porting a key word acting as the central player, which can be extracted by the function getCommu- nityCentrality(). However, the proper number of words to be extracted are not known in the present study.

Conclusion

The present paper proposes to further the develop- ment of a dictionary of classical Japanese poetry using pairwise term information which is generated by the community centrality procedure.

[2] Satsuki matsu / hana tachibana no / ka o kageba / mukashi no hito no / sode no ka zo suru of No. 13 in Chap- ter 3: Summer, the Kokinshu¯ (ca. 905) which appear in the Tales of Ise (ca. 800) as well.

We con- ducted an experiment using the R package “linked communities” and showed that the methods in the experiment extracted similar sub cluster terms which contribute to the description of classical Japanese poetry.

References

[1] Ahn, Yong-Yeol, James P Bagrow, and Sune Lehmann Jrgensen (2010) “Link communities reveal multiscale complexity in networks.”, Nature, Vol. 466, No. 7307, pp. 761–764.
[2] Csardi, Gabor and Tamas Nepusz (2006) “The igraph software package for complex network re- search”, InterJournal, Vol. Complex Systems, p. 1695.
[3] Freeman, Linton C. (1978) “Centrality in social networks conceptual clarification”, Social Networks, pp. 215–239.
[4] Hodoˇsˇcek, Bor and Hilofumi Yamamoto (2013) “Analysis and Application of Midrange Terms of Modern Japanese”, in Computer and Humanities 2013 Symposium Proceedings, No. 4, pp. 21–26.
[5] Kojima, Noriyuki and Eiz¯o Arai (1989) Kokin- wakashu¯, Vol. 5 of Shin-Nihon bungaku taikei (A new collection of Japanese literature), Tokyo: Iwanami shoten.
[6] Kubota, Jun (1979) Shinkokinwakashu¯, Shincho Ni- hon Koten Shu¯sei, Tokyo: Shinchosha.
[7] Nakamura, Yasuo, Yoshihiko Tachikawa, and Mayuko Sugita (1999) Kokubungaku kenkyu¯shiryo¯kan d¯etab¯esu koten korekushon (Database Collection by National Institute of Japanese Literature “Niju¯ichidaishu¯” the Sh¯oho edition CD-ROM): Iwanami Shoten.
[8] Shin-pen Kokkataikan Henshu¯Committee ed.(1996) Shimpen Kokka-taikan: CDROM Ver- sion: Kadokawa Shoten.
[9] Yamamoto, Hilofumi, Hajime Murai, and Bor Ho- doscek (2014) “Development of an Asymptotic Word Correspondence System between Classi- cal Japanese Poems and their Modern Translations”, in Proceedings of Computer and Human- ities 2014, Vol. 2014, pp. 157–162.

High-throughput Collation Workflow for the Digital Critique of Old Japanese Books Using Computer Vision Techniques

Asanobu Kitamoto (National Institute of Informatics), Kazuaki Yamamoto (National Institute of Japanese Literature)

Massive digital image collection of about 300,000 pre-modern Japanese books is expected to be released as open data in coming years thanks to the effort of the project "Building International Collaborative Research Network for Pre-modern Japanese Texts" lead by National Institute of Japanese Literature. One of the fundamental tasks in such a massive collection is collation, or more specifically, comparison of books to identify different editions and their relationship. Books with the same title may have different content, not only in terms of textual content, but also in terms of variants and impressions evidenced by small differences that are difficult to notice by human inspection. The goal of our research is to develop a high-throughput workflow for comparing different editions of books at the pixel level of digital images.

In contrast to text-based comparison, image-based comparison has advantages as follows. First, it does not require transcription of books before comparison. Second, it is also effective for non-textual comparison such as difference of paintings, or quality of printing, as long as books in comparison have the same layout with minor differences. Although text-based comparison is powerful to allow comparison beyond different physical layout, we believe that image-based comparison is relevant because this simple but tedious task is what computers can perform better than humans. This work, however, is still in a preliminary phase, and the following result is more of preliminary than comprehensive.

The whole workflow can be summarized as follows. First, a page divider tries to divide a digitized image into a set of page images for a page-to-page comparison. But the page divider heavily depends on specific capturing condition, so we can choose either automatic or manual approaches for this task. Second, using computer vision techniques, feature points are automatically extracted from page images of different editions. Extracting feature points is an active area of research in computer vision, and they generally give us satisfactory results. Please note, however, that an unsolved problem remains in comparison across images of different quality, such as full-color, gray-scale, and (nearly) binary images. Third, feature points are used as reference points for registration using rigid or non-rigid registration techniques. Rigid registration, which only involves shift, rotation and scale, usually gives satisfactory results for the purpose of inspecting minor difference, but non-rigid registration may be required for advanced analysis, such as local distortion of woodblock. Fourth, after registration, two images are superimposed and compared for each pixel to color-code intensity difference to highlight large difference. A useful color scale for a human inspector.is to assign red and blue color for large difference and white color for small difference.

Figure 1 shows a preliminary result about comparing two editions of the same book. The left panel shows the result of correspondence between reference points on two images. The right panel shows the color-coded difference between two editions after registration, illustrating that most of the pixels become white or gray due to cancelation of same characters on two editions. A human inspector can easily identify large differences in two editions represented by red or blue color, namely stamps in different locations.

Even if two editions are the same, however, two editions cannot be totally canceled to produce a purely white image due to following reasons. First, a page image contains not only characters but also other noises, such as stain on the paper, or partial transparency of the paper showing characters on the other side. Second, local variation cannot be removed by a simple rigid registration, such as local distortion of the woodblock at the edge, or intensity variation of the ink in the middle. A human inspector, however, can quickly filter out those noises, and can easily identify meaningful differences without influence of subjectivity in human reading.

A future work is to build an edition comparison service for comprehensive image-based analysis of book editions. When an image of one edition is uploaded to the service, the server compares the uploaded image with other editions in the storage, and suggest that it is one of the existing

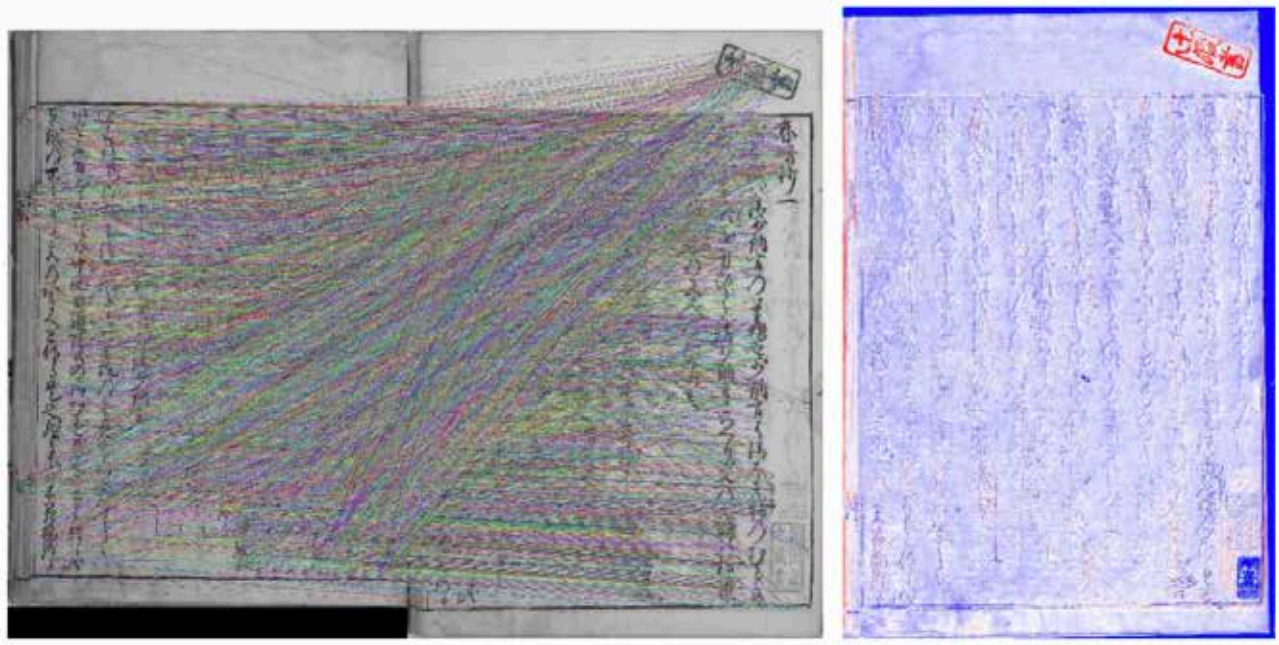

Figure 1: Matching two images using reference points extracted from two images, and the comparison of two images using red/white/blue color scale.

editions or is a new one. This may be a killer app for the archive of old Japanese books because having more editions, variants, and impressions in the storage means higher accuracy of comparison, which is the reason to attract more users. This kind of positive feedback is known as network effect.

Lastly, we would like to emphasize that the target of this research is at the level of text critique, but not at the level of text interpretation. This is one example of our proposed concept "digital critique" which uses information technology to enhance a traditional human-based criticism. We expect that this workflow is beneficial to scholars because it will reduce the burden of scholars who need to perform a tedious text critique task of character-by-character comparison, and it will allow them to focus more on a higher level of research such as text interpretation.

Acknowledgment

The project was supported by collaborative research grant from National Institute of Japanese Literature. Registration is performed using open source software, OpenCV. The books used in the experiment is (1) 枕草子春曙抄, 国文研高乗, and (2) 春曙抄, 国文研鵜飼.

Development of Glyph Image Corpus for Studies of Writing System

Yifan Wang (University of Tokyo)

We have built a software suite to auto-generate, edit, and annotate glyph image databases in order to serve our text / glyph image integrated corpus of dictionaries Yiqiejing Yinyi (一切經音義) and Xu Yiqiejing Yinyi (續一切經音義) in a printed Chinese Buddhist canon Taishō Tripiṭaka (大正新脩大藏經).

The software has three main components. 1) Character isolation system (fig. 1), which automatically detects and crops each character from digital facsimiles of the books. The program has processed all input images with approx. 94% accuracy, where existing commercial OCR programs failed to correctly detect vertical lines and/or warichu style (inserting in-line annotation in double lines of smaller size characters) layout. 2) Glyph image editor (fig. 2), which has mainly been used to correct auto-generated character coordinates output by the isolation system. The program allows users to visually browse each page and quickly find errors. 3) Glyph comparison and annotation interface (fig. 3), that runs as web application, and on which users can search a certain character to compare all (or some of) appearances en masse in images stored in the corpus. It is also designed to quickly add metadata to correctly categorize glyphs into each group that consists of those regarded as the same shape. All aforementioned programs, including the corpus itself are built upon open-source libraries (OpenCV, Qt, Ruby etc.), thus easily customizable according real use cases. They, as well as their dependencies, also maintain high portability, being functional in all Windows, Mac OS X, and Linux platforms. The programs enabled us to reduce considerable amount of time and manual work, efficiently develop the corpus, and continuously maintain and improve the data set without expert knowledge in computing.

The corpus is focused on analyzing and obtaining statistical data on the internal graphemic system (i.e. whether two distinct glyphs are considered same in quality) in those documents, and consists of text data derive from SAT Project (providing digitalized text of Taishō Tripiṭaka) and the generated glyph DB. Yiqiejing Yinyi and Xu Yiqiejing Yinyi in Taishō Tripiṭaka show unique features even compared with other parts of the collection. Despite the fact that the tripiṭaka is a letterpress printing, they embrace a vast number of character variants; est. 30,000 different glyph types of varied degree of similarity are recognized, with approx. 3,000 characters are preliminary found to be subject of addition in the Unicode character set, roughly as many as the number we proposed to Unicode from all other portions of the publication. This exceptional diversity is accounted for by complicated aspects such as their fidelity to Tang-dynasty handwriting convention, multiple references with mixed collation history used during edition, and interaction of them with modern interpretation and possibly technical errors in editing. As we are preparing for Unicode proposal to encode characters in Taishō Tripiṭaka, it is urgently needed to understand the structure of the entangled writing system from the sections, which contain over 1,000,000 characters in total, hence difficult for small group of researchers to conduct an exhaustive analysis. And this is the reason we introduced automatic processing.

As we are now working on accurate glyph categorization using the programs, we will share some of our findings at the conference in September.

We believe that the system we use is also applicable to other grammatological or philological studies that require fine-grained analysis of each single character and use printed East Asian documents with vertical layout as materials.

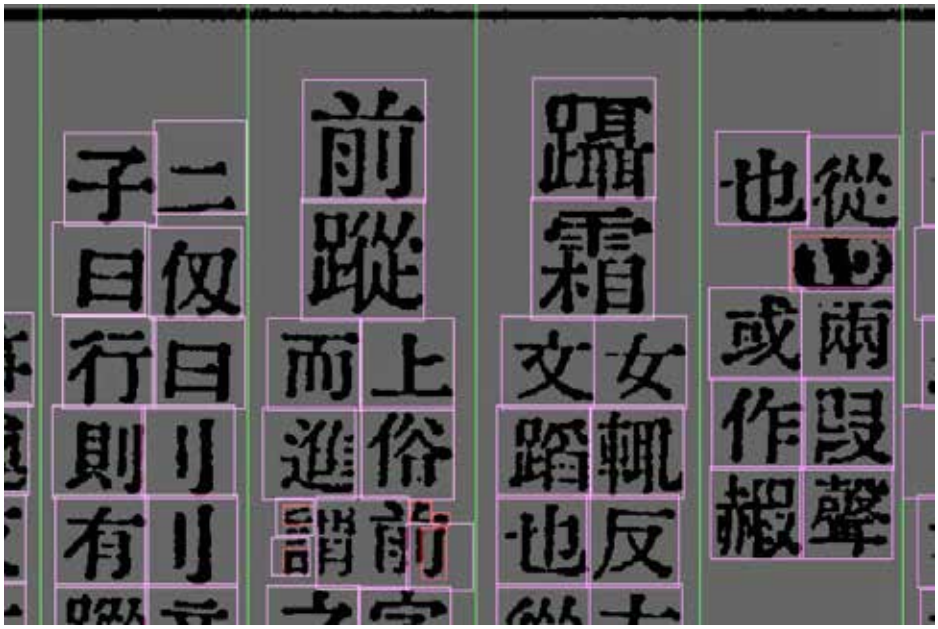

Figure 1:

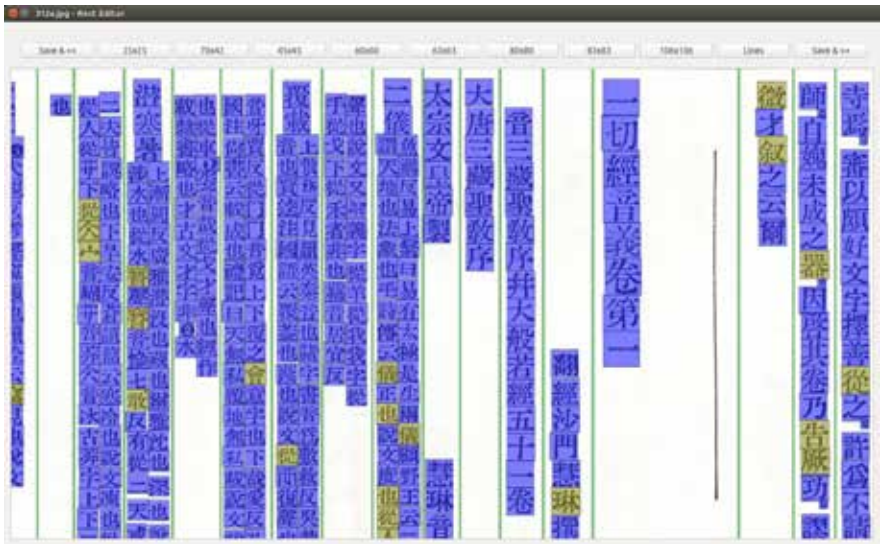

Figure 2:

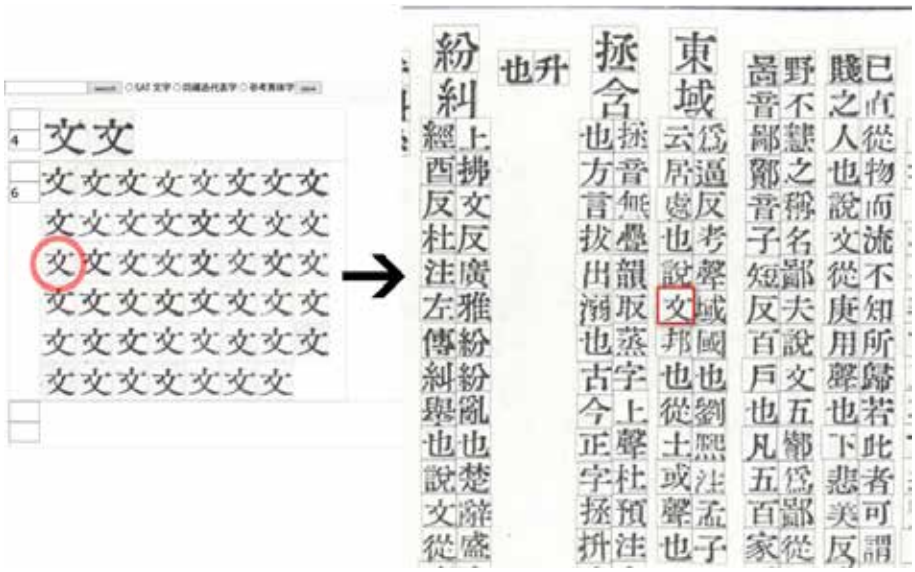

Figure 3:

Relationship between film information and audience measurement at a film festival

Masashi Inoue (Yamagata University)

Abstract
This paper presents the results of an analysis on the relationship between film information and audience measurement at a film festival. The aim of the analysis is to create a model that can predict attendance at the halls and the congestion rate of halls and identify the important attributes at screenings. The results of the analysis revealed that the categorization of films screened is the most important factor for the audience to attend film screenings.

Introduction

Artistic contents are delivered to audiences more often in digital format via digital networks. In stark contrast to convenient consumption through digital transmission, live performances are sometimes considered a better way to fully enjoy artistic content, a notion that has gained popularity in recent times. When the contents are films, film festivals are considered a form of live performance (Bordwell, Thompson, & Ashton, 2004). During the festival, both the creators and the audiences get together and discuss the films that are screened. Until now, little has been known about film festivals as a media beyond the publically known festival organization and the official statistics provided by the organizers. Exceptions are the analysis of film selection processes in a film festival (Inoue & Sakuma, 2014) and the special journal issue focusing on the historical and geographical diversities in film festivals (Papadimitriou & Ruoff, 2016). The current work is an attempt to understand the properties of film festivals in terms of audience participation by building prediction models of hall attendance and congestion rates. A similar attempt has been made to predict box-office revenues from the search statistics on upcoming films (Google, 2013). However, compared with major commercial films, artistic films shown in a film festival have little information on the potential audiences. Therefore, we focused on the information about the films and the organization of the film festival for building the prediction model.

Data

We considered the Yamagata International Documentary Film Festival (YIDFF) as the target event. YIDFF is held biennially. We used data from the 174 films screened in the year 2011. The information about the films were either provided by the organizers or retrieved by Web crawling on the festival website.

Method

We used multiple regression and random forest to construct the prediction models. These methods were chosen prioritizing interpretability over accuracy of prediction. The dependent variables were either the raw data on the number of the audience or the congestion rate of the hall. We mainly discuss the congestion rate model here. The independent variables were as follows: number of countries involved in film production (real number), running time (real number), capacity of the halls (real number), talk held after screening (binary), weekday or holiday (binary), number of films the director appeared in previous YIDFFs (real number), program (one of 8 categories), starting time (real number), and the number of audience in the previous film in the same hall (real number). The 8 programs considered are as follows: IC (International Competition: 15 outstanding films selected from entries from around the world); NAC (New Asian Currents: Introducing up-and-coming Asian documentary filmmakers); NDJ (New Docs Japan: A selection of new Japanese documentaries); IS (Islands/I Lands, NOW—Vista de Cuba: A program focusing on Cuba as an "Island"); MT (My Television: A program featuring Japanese TV documentaries, with a focus on works from the 1960s and 1970s); TJ (A Reunion of Taiwan and Japanese Filmmakers: 12 Years

Later: Filmmakers from YIDFF New Asian Currents '99 return with old and new films); FY: (Films about Yamagata: The third edition of this regular program that looks at Yamagata and its relation to cinema); CU (Great East Japan Earthquake Recovery Support Screening Project "Cinema with Us").

Result

When multiple regression analysis was used, the adjusted coefficient of determination was found to be 0.43. When random forest was used, the adjusted coefficient of determination was 0.37. Both values are lesser than 0.5, which is often the threshold for reliability. Therefore, we could not obtain a reliable prediction model from the available data. The factors contributing to the prediction of congestion rates were the capacities of the halls (as per the regression analysis) and the programs (as per both methods).

Conclusion

We analyzed film popularity based on audience measurement in the Yamagata International Documentary Film Festival (YIDFF). This analysis based on multiple regression and random forest methods indicated that the programs as part of which the films are screened are an important factor for predicting higher audience participation. For example, the organizers had assigned halls with similar capacities to two special programs: CU (Great East Japan Earthquake) and IS (Cuba). However, the program CU had more audience participation than the program IS, probably because the audiences were more attracted to a familiar and current topic.

Acknowledgements
This work is based on an analysis performed by Yuri Koseki. Kazunori Honda helped to improve this abstract.

References
[1] Bordwell, D., Thompson, K., & Ashton, J. (2004). Film art: An introduction (7 ed.). New York: McGraw-Hill.
[2] Google. (2013, 6). Quantifying Movie Magic with Google Search.
[3] Inoue, M., & Sakuma, S. (2014). Analysis of the film selection process for a film festival. The 7th International Workshop on Information Technology for Innovative Services (ITIS-2014), (pp. 582- 587). Victoria, Canada.
[4] Papadimitriou, L., & Ruoff, J. (2016). Film festivals: origins and trajectories. New Review of Film and Television Studies, 14 (1), 1-4.

Linking Scholars and Semantics: Developing Scholar-Supportive Data Structures for Digital Dūnhuáng

Jacob Jett, J. Stephen Downie (University of Illinois at Urbana-Champaign), Xiaoguang Wang (Wuhan University), Jian Wu, Tianxiu Yu (Dunhuang Research Digital Center), Shenping Xia (Dunhuang Research Academy)

Introduction

The Digital Dūnhuáng Project (Wu, 2015; Zhou 2015) is a very large-scale field digitization project in the process of digitizing the contents of the Mògāo Caves, Dūnhuáng's vast system of 492 Buddhist temples and cave sites. The caves contain thousands of sculptures, murals, and other cultural artifacts that were fashioned during the thousand years (~400-1400 CE) that the city served as a crossroads on the Silk Road and vital Buddhist cultural center. The Mògāo Caves are a UNESCO World Heritage Site and are of interest to both scholars and the general public alike. The level of interest in this cultural treasure is reflected by the 1.1 million visitors to the caves in 2015 alone.

There has been a great deal of effort, realized through the International Dūnhuáng Project 1 (IDP), to digitally preserve and publish the many manuscripts found in Cave 17. More recently, the Digital Dūnhuáng project of the Dūnhuáng Academy has been digitally capturing the sculptures, paintings, and other important cultural artifacts found within the caves. They are creating high resolution images so that they may be made more accessible to scholars worldwide and shared with those unable to physically travel to Dūnhuáng (Wang, 2015). Thus far the project has only digitized the contents of 120 of the 492 caves. Despite the modest number of caves photographed, the Digital Dūnhuáng project has already produced 941,421 digital images of the cultural artifacts. We estimate that by the project's end, almost four million digital images will have been produced.

Digital Infrastructure

In this poster abstract, we present a proposed formal metadata model designed to improve the utility of the soon-to-be millions of Dūnhuáng cave images with the special intention of enhancing the impact of these important resources on digital and traditional humanities and religion scholarship worldwide. The process of digitization—the production of digital photographs—of the Mògāo Caves rich repository of cultural heritage is an ongoing process.

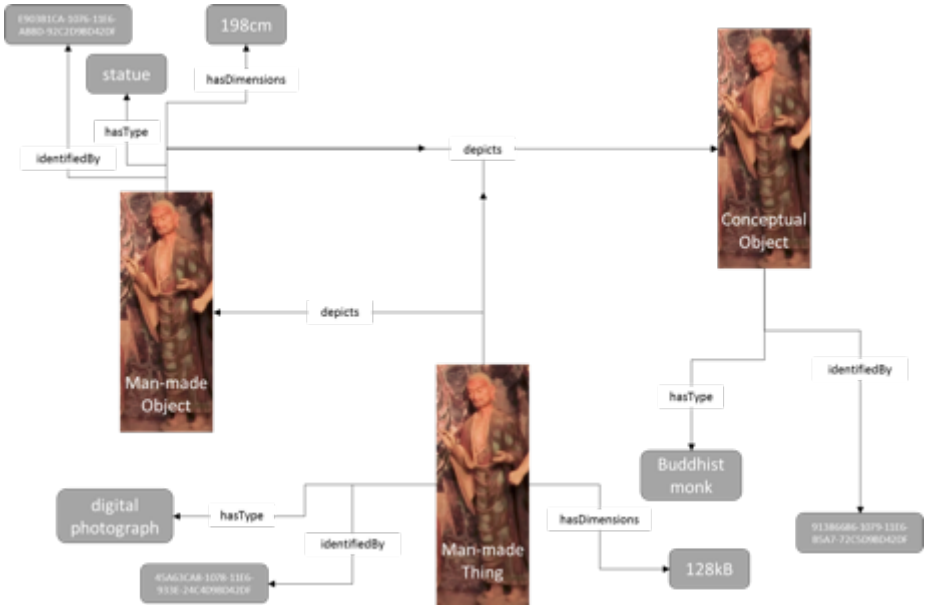

Figure 1. Persistent identifiers and base taxonomic classification

We assert that the digital annotation of the Dūnhuáng photographs and the things denoted in them is a key aspect for providing remote scholars the means to interact with this treasure trove of historic works. Thus, before any digital annotation can take place, we propose that a necessary first step is to inventory and identify the cultural artifacts in the caves (Downie, 2015). Figure 1 (above) illustrates one method in which this can be done, creating a rich interlinked web of man-made objects and the conceptual objects they depict.

The creation of persistent identifiers for all of the caves' contents at their various intellectual levels of scholarly interest is the cornerstone upon which our proposed interactive digital infrastructure is to be built. Once an inventory of persistent, web-accessible objects has been put into place, then scholars may interact with the various intellectual targets for scholarship by adding their own unique layers of digital annotations.

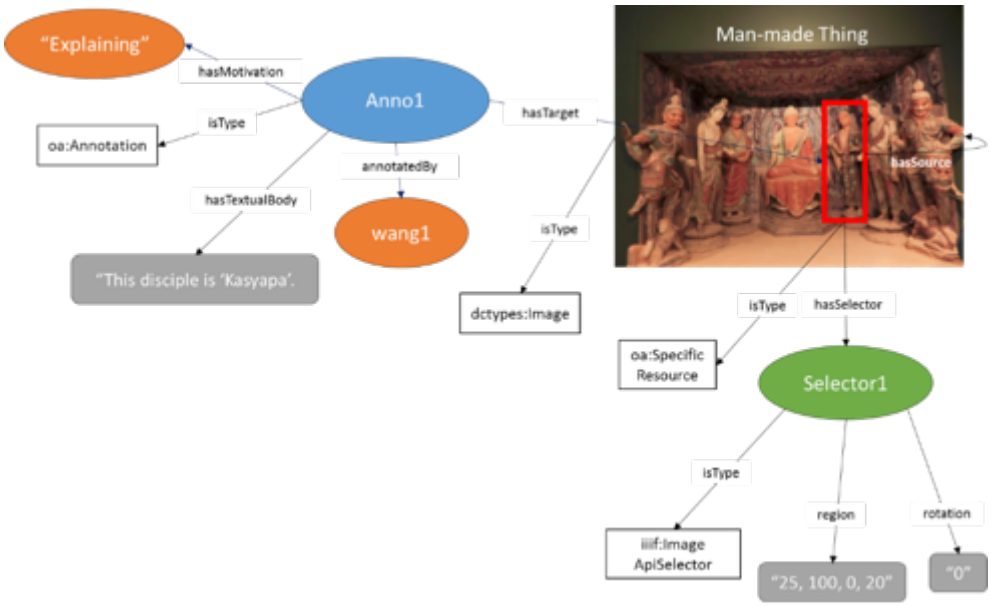

Figure 2. Simple scholarly annotation[1]

As Wang et al. (2016) observe, metadata, deep semantic analysis and topical indexing are among the kinds of annotation taking place with regards to the digital photographs being produced by Digital Dūnhuáng. Figure 2 (above) illustrates a simple scholarly annotation scenario. In this example, a scholar has labeled the target conceptual object (the disciple) in the red box with a name, "Kaspaya."

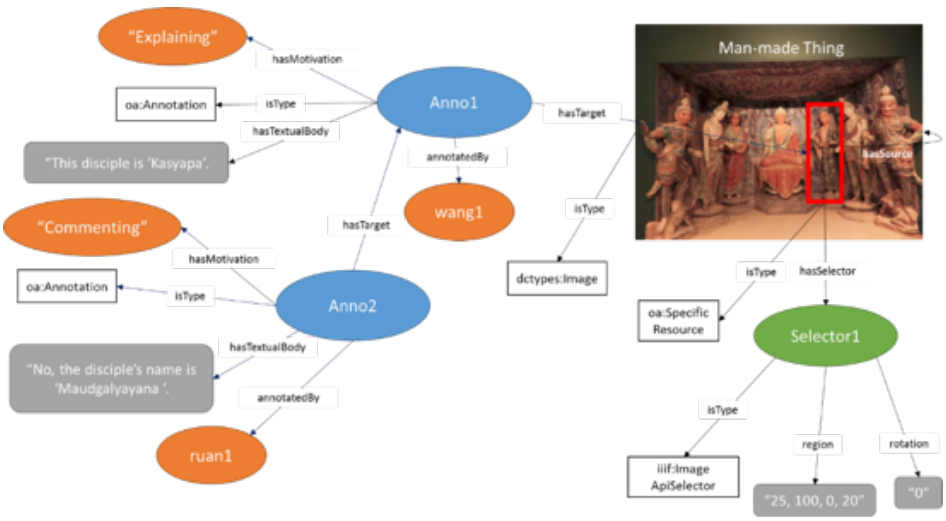

Figure 3. Direct scholarly discourse through digital annotation

[1] Note that for the sake of readability, many core annotation properties concerning the annotations' provenance, such as date created, have been left out of these illustrated examples. The annotation model's full property set can be found at: https://www.w3.org/TR/annotation-model/

These technologies make use of linked data (Berners-Lee, 2006; Bizer et al., 2009) through RDF[2]-conformant ontologies and serialization formats, such as JSON-LD[3]. Once a digital foundation of persistent identifiers and basic categorization has occurred and annotation infrastructure has been implemented, the scholars may interact directly or indirectly with one another through the act of annotating (illustrated in Figures 3 (above) and 4 (below)). In this example, a second scholar adds a dissenting view of what the disciple's name should be, saying "no, this disciple's name is 'Maudgalyayana'."

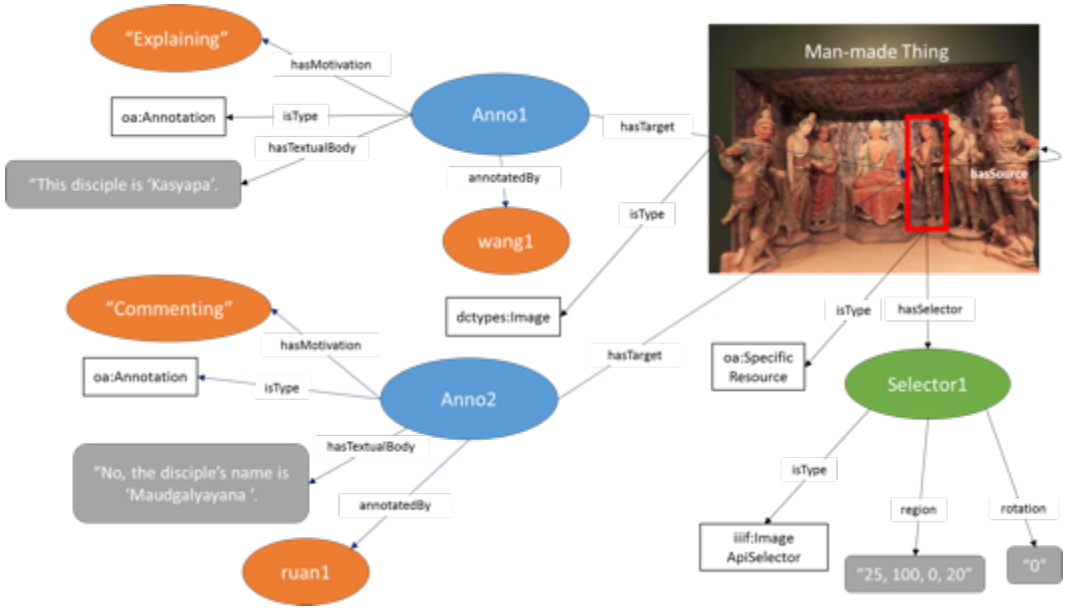

Figure 4. Indirect scholarly discourse through digital annotation

These illustrative examples merely showcase one of the many scholarly discourse use cases—promoting discourse—digital annotations of this kind can play. These annotations may also be part of a process for arriving at a consensus for the identity of the monk depicted by the statue or they might record a narrative of discussions about the caves' contents. Digital annotations like these might also be applied in classroom settings, permitting students and instructors with means to interact with the cultural objects that they would not normally have.

Of course, the mechanics and limitations of digital systems are such that it is not always apparent that the annotators are actually naming the same entity. As Arms (1995) observes, the scholarly users of the Digital Dūnhuáng's images do not want to interact with the digital photographs as much as they would like to make assertions regarding the things denoted within the photographs. One potential method for remedying this problem is to extend the framework with properties that are designed to operate in parallel to process of anchoring annotations to their targets. An example of this appears in Jett et al. (2016) and is illustrated in Figure 5 (below).

In this case the property, "hasTargetFocus" is used to preserve the fact that the two scholars are discussing the same abstract thing, the old disciple, even though their annotations are anchored to two completely different entities (i.e., to a region of a photograph and to an annotation of the region of that photograph, respectively). This level of representation is useful even if their annotations where anchored to precisely the same target because it clarifies that their annotations are about the monk depicted by the statue and not the statue itself or the photograph that depicts it.

Another advantage that digital knowledge representation systems bring is the flexibility of extensible frameworks. Not only do extensible frameworks allow more of a scholar's intentions to be preserved they also permit choice of domain vocabularies for description of resources (e.g., CIDOC-CRM[4]) and the ability to support specialized digital tools. For example, scholars using Digital

[2] https://www.w3.org/RDF/
[3] http://json-ld.org/
[4] http://www.cidoc-crm.org/html/5.0.4/cidoc-crm.html

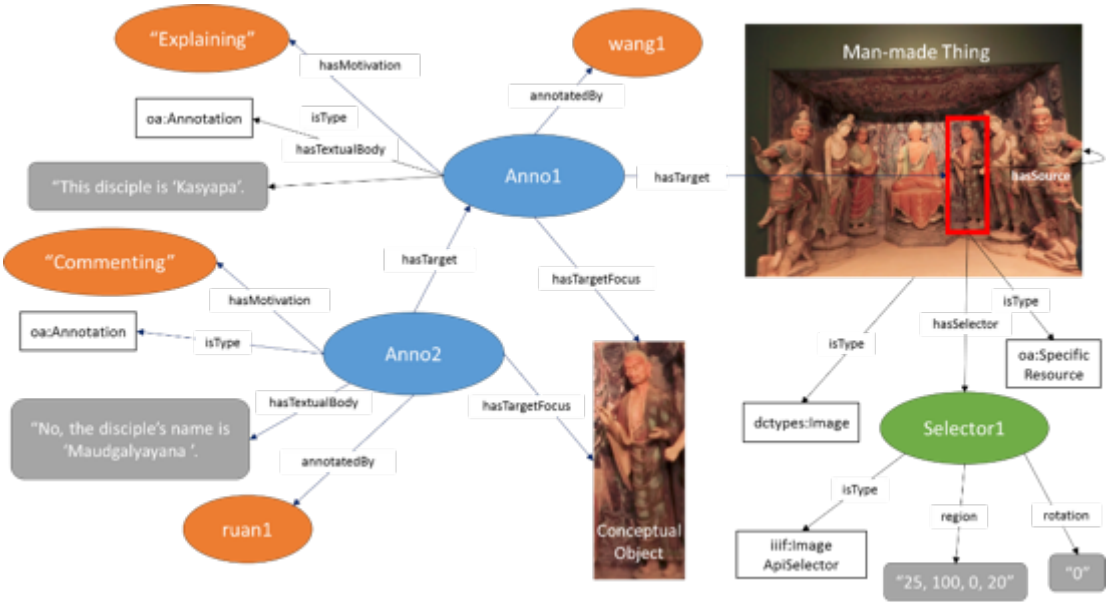

Figure 5. Preserving the intellectual focus of scholarly discourse

Dūnhuáng might wish to use the International Image Interoperability Framework's image selector[5], which allows them to rotate the subject of an image in three dimensions as well as specifying some particular part of an image. Similarly, the use of this framework, will allow scholars to gather up all of the annotated instances of, for example, the disciple "Kaspaya" from all of the Dūnhuáng caves across time and space. Persistent identifiers and a basic categorical framework are the cornerstone for building a digital scholarly workplace.

References

[1] Arms, W. Y. (1995). Key concepts in the architecture of the digital library. D-Lib Magazine 1(1). Available via: http://www.dlib.org/dlib/July95/07arms.html

[2] Berners-Lee, T. (2006). Linked data. Designed Issues: Architectural and Philosophical Points. Accessible via: https://www.w3.org/DesignIssues/LinkedData.html

[3] Bizer, C., Heath, T. & Berners-Lee, T. (2009). Linked data—The story so far. International Journal on Semantic Web and Information Systems 5(3), pp 1-22. DOI: 10.4018/jswis.2009081901

[4] Downie, J. S. (2015). "Enhancing the impact of Digital Dunhuang on digital humanities scholarship." Panel presentation given at DH 2015 (Sydney, Autralia, 30 June – 3 July 2015).

[5] Jett, J., Cole, T. W., Dubin, D. & Renear, A. H. (under review). "Discerning the intellectual focus of annotations." Paper submitted to Balisage: The Markup Conference 2016 (North Bethesda, MD, 2-5 August 2016).

[6] Wang, E. (2015). "Explicating the potentials of Digital Dunhuang on scholarship and teaching." Panel presentation given at DH 2015 (Sydney, Autralia, 30 June – 3 July 2015).

[7] Wang, X., Song, N., Zhang, L., Jiang, Y. & Marcia, Z. (2016). Understanding the subject hierarchies and structures contained in Dunhuang murals for deep semantic annotation: A content analysis. Unpublished working paper to be submitted.

[8] Wu, J. (2015). "Introducing the 'real' Dunhuang and the Digital Dunhuang project." Panel presentation given at DH 2015 (Sydney, Autralia, 30 June – 3 July 2015).

[9] Zhou, P. (2015). "Digital Dunhuang: Digitally capturing, preserving, and enhancing real Dunhuang." Panel presentation given at DH 2015 (Sydney, Autralia, 30 June – 3 July 2015).

[5] http://iiif.io/api/annex/openannotation/#status-of-this-document

A Web Based Service to Retrieve Handwritten Character Pattern Images on Japanese Historical Documents

Akihito Kitadai (J. F. Oberlin University), Yuichi Takata, Miyuki Inoue, Guohua Fang, Hajime Baba, Akihiro Watanabe (Nara National Research Institute for Cultural Properties), Satoshi Inoue (University of Tokyo)

We present a web based service to retrieve handwritten character pattern images written on historical Japanese documents.

Digital images of handwritten character patterns are important research products of history and archaeology. We have been providing two digital archives of the images. One of them contains the images extracted from mokkans written in and around 8th century. The mokkan is a Japanese name of a type of historical documents. Wooden tablets were used as the recording media, and brushes with Indian ink were used to write the character patterns of the documents. The other contains the images from paper documents written in and around 9-18th century. Every image of the character pattern is selected by experts of Japanese history, archaeology and calligraphy.

Information retrieval methods and technologies are critical factors for digital archives of history and archaeology. Employing a character code as a key of the retrieval is a reasonable implementation for digital archives of character pattern images. We are providing a crossover retrieval system of the two digital archives in which both the archives output the images that belong to the key code (http://r-jiten.nabunken.go.jp/kensaku.php). However, the character codes for historical languages have not been defined clearly yet. The definitions are ongoing research activities of history and archaeology. For the reason, we need to provide alternative methods that employ other information as the retrieval key.

The web based service Mojizo that we present in this abstract is one of the alternatives. As same as our system previously mentioned, Mojizo provides cross over retrieval of the two digital archives, but it employs a handwritten character pattern image as the key.

Mojizo has a shape evaluation engine consisting of pattern matching technologies. This engine calculates similarity between the key and the images on the digital archives. Since the evaluation needs a large amount of calculation, we designed and implemented the engine and the other modules of Mojizo to run on server side. Therefore, we can use Mojizo via small portable terminal devices with network connection and low computing power only. Digital cameras commonly equipped on such portable terminal devices work well to capture the key images of handwritten character patterns on historical documents. We have opened Mojizo on our web site (http://mojizo.nabunken.go.jp/). Web browsers provide user interfaces to input the key images and to see the similar handwritten character pattern images. Mojizo also provides the links to meta data sets for each of the similar images. The meta data sets are results of decoding processes of historical documents performed by historians and archaeologists. Therefore, we expect that Mojizo supports users who have unreadable handwritten character pattern images.

To broaden application ranges of the digital archives is an aim of our research activities. The users of Mojizo need no keyboard to input the character codes. This means that Mojizo can provide ubiquitous gateways to the digital archives. Activating usage of digital archives is important to inherit the history of the human behavior in our modern society. Mojizo is providing about 28,000 images of handwritten character pattern with the link to their meta data sets, and the number is increasing.

Our presentation will display the detail design and implementation of our web based service including the shape evaluation engine. Also, we will present some examples of information retrieval using Mojizo.

Acknowledgment

This work was supported by the Grants-in-Aid for Scientific Research (S)-25220401, (A)-26244041 and (C)-15K02841.

Image recognition and statistical analysis of the Gutenberg's 42-line Bible types

Mari Agata (Keio University), Teru Agata (Asia University)

Traditionally, analyses of types used in the early printed books have been conducted by naked but trained eyes of bibliographers. The types of the Gutenberg 42-line Bible (hereafter "B42"), the earliest printed book in Europe with movable metal type, is no exception.

In 1900 Paul Schwenke published results of his minute and painstaking investigation of the B42 type.[1] He identified and listed two hundred ninety types. The reasons for such a large number of types are the existence of abbreviations, contractions, and secondary forms, or abutting types, of almost every letter of the alphabet. The left side of an abutting type was flat, without the diamond shaped spur, so it could be placed close to the preceding type according to defined rules. Schwenke observed that after letters c, e, f, g, r, t, x, and y, this abutting type was used.

This composition rule was so strict that some deviations were even corrected during the actual print run, as the collation using superimposition of digital images by the present author demonstrated.[2]The collation also raised new questions about the composition rules. For example, four stop-press corrections concern a shorter abutting "r"; its usage has not been previously studied in detail and thus need further analysis. In addition, collation results suggest that the types were not perfectly locked up but set loosely, resulting in many variations of word spacing, shifted lines, and both inclined and drifted letters.

Furthermore, other scholars had identified a different number of types of B42. Schwenke's close observation may require several amendments.

In 2000, Paul Needham and Blaise Agüera y Arcas questioned how Gutenberg cast his types.[3] A traditional view is that he produced types by steel punch, copper matrix, and adjustable hand mould, and thus he could produce thousands of "identical" types, from a single matrix. Needham and Agüera y Arcas made a clustering analysis of the lower case "i"s used in a 20-page Papal Bull printed in the DK type, which was made earlier than the B42 types and closely resemble to them. Several hundred "i" clusters were discovered; a far greater number than expected. They claimed that these "i" types could not have been made from a common punch and matrix and suggested that many matrices had been used in parallel, or equivalently, the matrix had been temporary and needed to be re-formed between castings. This is a significant question to shake to the foundations of the printing history. In spite of the considerable attention their research attracted, there have been few substantial follow-up studies.[4]

The adoption of computer-based research now allows us to conduct experiments on a much larger scale that was previously possible. The present authors have developed a new method of semi-automatic image recognition of the B42 types and demonstrated that it have explanatory power beyond the influence of inking and photographic conditions when applying to data of a large scale.[5]

---

[1] Paul Schwenke, Untersuchungen zur Geschichte des ersten Buchdrucks. Berlin, Behrend, 1900.

[2] Author. Stop-press Variants in the Gutenberg Bible: The first report of the collation. The Papers of the Bibliographical Society of America. 2003, vol. 97, no. 2, p. 139-165; Author. デジタル書物学事始め：グーテンベルク聖書とその周辺. 勉誠出版, 2010 [Author. Introduction to digital bibliography: the Gutenberg Bible and beyond. Bensei Shuppan, 2010.]; Author. "Improvements, corrections, and changes in the Gutenberg Bible." Scribes, Printers, and the Accidentals of their Texts. Frankfurt am Main, Peter Lang, 2011, p. 135-155.

[3] Agüera y Arcas, Blaise. "Temporary Matrices and Elemental Punches in Gutenberg's DK Type." Incunabula and Their Readers: Printing, Selling and Using Books in the Fifteenth Century, Jensen, Kristian, ed. London, British Library, 2003, p. 1-12.

[4] Pratt, Stephen. The myth of identical types: A study of printing variations from handcast Gutenberg type. Journal of the Printing Historical Society. 2003, new series 6, p. 7-17.

[5] Authors. 活字の識別とその応用：グーテンベルク聖書の活字のクラスタリング. 日本図書館情報学会 2014 年度研究大会. 2014-11-29, 梅花女子大学（大阪府）. 第 62 回日本図書館情報学会研究大会発表論文集. 2014, p. 117-120 [Authors. Recognition of types and its bibliographical application. Annual conference of Japan Library and Information Science. 2014-11-29, Baika Women's University.]; Authors. A newapproach to image recognition

The purpose of this study is to make further analysis of the B42 types with an improved method of image recognition reinforced by machine learning. The image data of B42 held in the Keio Gijuku Library was used for analysis. Information about X and Y coordinates, pixel width and height, and transcribed characters of each type image data are collected and used for the statistical analysis.

To analyze the vertical alignments, the average variance of the Y coordinate for each type image of each line, excluding types with descenders and capitals, were calculated. When doing a page-by-page variance analysis, pages that were thought to have been printed earlier exhibited greater variance.

The width data of each type image provided us useful information. A frequency distribution of the width of several types had two mild peaks; the wider types were those of primary forms, while the more narrow ones were those of secondary, abutting forms. Transcribed character data showed that the narrower ones positioned after letters c, e, f, g, r, t, x, and y. This result supports one of the composition rules observed in Schwenke's study.

Further statistical analyses enable to investigate such characteristics as variance in the body size, the relative distance between a contraction bar and a main letter, and more. A close examination of these characteristics will lead to identify type variants and their distribution in the book. An accumulation of the results could give further clues to questions regarding specific details of the first printing shop in Europe, and, hopefully, of Gutenberg's casting method.

and clustering of the Gutenberg's B42 types. Memory, the (Re-)Creation of Past and Digital Humanities.-2016-03-15, Keio University (Tokyo).

Comparisons of Different Configurations for Image Colorization of Cultural Images Using a Pre-trained Convolutional Neural Network

Tung Nguyen, Ruck Thawonmas, Keiko Suzuki, Masaaki Kidachi
(Ritsumeikan University)

Introduction

This paper describes image colorization of cultural images, such as ukiyo-e, by which colors are added to grayscale images. This is done in order to make them more aesthetically appealing, culturally meaningful, or even inspiring. Importance of this task can be seen, for example, by a relatively large portion of grayscale images in the archive portal of the Art Research Center (ARC), Ritsumeikan University, e.g., 1600 grayscale images out of 4588 images of the type Yakusha-e (actor painting) publicly accessible.
In this work, we followed the same approach as Gatys et al. [1] that uses a pre-trained convolutional neural network (CNN), called VGG-19 [2], for transferring the style of an image to another image while maintaining the content of the latter one. In particular, using ukiyo-e images from the aforementioned archive, we investigated a number of configurations for setting VGG-19's layers, weighting between the style loss and the content loss, and optimizing the parameters. Discussions are done that give insights to future work.

Methodology

The content of a grayscale image is combined with the style of a color image, resulting in colorizing the grayscale image. For a layer l in the network, we denote the number of feature maps and the size of each feature map in that layer as N_l and M_l, respectively. The content loss is then calculated by

$$L_{content}(p,x) = \frac{1}{N_l M_l} \sum_{i,j} \left(P_{ij}^l - F_{ij}^l\right)^2,$$

where $P^l \in \mathbb{R}^{N_l \times M_l}$ and $F^l \in \mathbb{R}^{N_l \times M_l}$ are the content representations, i.e. the features, of the content image p and the output image x, respectively. On the other hand, the style representation at layer l is given by the Gram matrix $G^l \in \mathbb{R}^{N_l \times M_l}$:

$$G^l = F^l (F^l)^T$$

and the style loss at layer l is calculated by

$$E_l = \frac{1}{N_l^2} \sum_{i,j} \left(\frac{A_{ij}^l}{M_l} - \frac{G_{ij}^l}{M_l} \right),$$

where A^l and G^l are the style representations of the style image aa and the output image x, respectively. Then the style loss function is defined considering style losses at multiple layers:

$$L_{style}(a,x) = \sum_l w_l E_l,$$

where w_l is the weighting factor of E_l, and equals to one divided by the number of layers. In addition, to smoothen the output image, we make use of the total variation regularizer given below:

$$L_{tv}(p,a,x) = \sum_{i,j} \left(\left(x_{i+1,j} - x_{i,j}\right) \right)^2$$

Finally, the total loss is calculated as the weighted average of the aforementioned losses:

$$L(p,a,x) = \alpha L_{content}(p,x) + \beta L_{style}(a,x) + \gamma L_{tv}(x).$$

Evaluation

We conducted various experiments to compare different configurations. Two layer settings by Gatys et al. [1, 3] and one by Yin [4] were considered. We also compared the use of stochastic gradient descent (SGD) [1] with that of LBFGS [3], as an optimization algorithm for finding a

Table 1. Description of configuration names.

Index	Value	Meaning
X	1	Content layer: conv4_2 Style layers: conv1_1, conv2_1, con3_1, conv4_1, conv5_1 [1]
	2	Content layer: conv4_2 Style layers: conv1_1, pool1, conv2_1, pool2, conv3_1, pool3, conv4_1, pool4 [3]
	3	Content layer: conv5_2 Style layers: conv1_1, pool1, conv2_1, pool2, conv3_1, pool3, conv4_1, pool4, conv5_1 [4]
Y	1	$\beta/\alpha = 0.001$
	2	$\beta/\alpha = 0.01$
	3	$\beta/\alpha = 0.1$
Z	1	L-BFGS with fixed β/α
	2	SGD with fixed β/α
	3	L-BFGS with decreasing β/α
	4	SGD with decreasing β/α

Figure 1. Relative error of each configuration at the last iteration.

minimum of LL. Moreover, we investigated the effect of decreasing β/α by 0.25% after each iteration [5].

Combing the aforementioned layer settings, optimization methods and different values of β/α leads to 36 different configurations in total. We use the format configX_Y_Z for naming each configuration; the value and meaning of each index X, Y, Z are provided in Table 1. For each configuration, we performed colorization 100 times, combining each content of 10 grayscale images with each style of 10 color images. γ was set to 0.001 and the output image was initialized with the content image as done in [5].

Because the range of the total loss varies considerably depending on the set of layers, we instead used the relative error defined below as a metric to compare different configurations:

$$Err(p,a,x) = Err_{content}(p,x) + Err_{style}(a,x),$$

where $Err_{content}$ and Err_{style} are the relative content error and the relative style error respectively.

$$Err_{content}(p,x) = \sum_{i,j} \frac{|P_{ij}^l - F_{ij}^l|}{(|P_{ij}^l| + |F_{ij}^l|)/2}$$

$$Err_{style}(a,x) = \sum_{l} W_l \sum_{i,j} \frac{|A_{ij}^l - G_{ij}^l|}{(|A_{ij}^l| + |G_{ij}^l|)/2}$$

Figure 1 shows the relative error of each configuration at the last iteration of the colorization process. The best configurations in terms of relative content error, relative style error and relative error are config3_1_4, config2_3_1, and config3_3_1 respectively. However, because from cultural viewpoints, it is important to maintain the original content, we visually compared images generated

Figure 2. Sample images generated with different configurations: config3_1_4 (the smallest relative content error), config2_3_1 (the smallest relative style error), config3_3_1 (the smallest relative error), config3_3_4 (the visual best).

by the best five configurations in terms of the relative content error and finally selected config3_3_4 as the visual best.

Figure 2 show results for three typical content-style pairs. Both config2_3_1 are config3_3_1 are colorful, but texture information in the content image, in particular kimono patterns and Japanese letters, is distorted. Visual comparisions reveal that while not as colorful as config2_3_1 or config3_3_1, config3_3_4 best preserves the original content while having color essence of the style image. As this kind of colorization is not only aesthetically appearing but also culturally meaningful, it may inspire the CNN user's artistic creativity.

Conclusions

In this paper, we compared a number of configurations for colorizing grayscale images by utilizing a pre-trained CNN, VGG-19. Our finding is that config3_3_4 leads to the best visual results regarding both original content preservation and color essence introduction. This work was based on existing work using VGG-19 for style transfer, not specifically for color transfer. In the future, we plan to develop loss functions and other settings that can ignore texture information and transfer only color information from the style image while preserving the content in the original image.

References

[1] Gatys, L.A., Ecker, A.S. and Bethge, M., 2015. A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576.

[2] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. Presented at International Conference on Learning Representations 2015.

[3] Gatys, L., Ecker, A.S. and Bethge, M., 2015. Texture synthesis using convolutional neural networks. In Advances in Neural Information Processing Systems (pp. 262-270).

[4] R. Yin, 2016. Content Aware Neural Style Transfer. arXiv preprint arXiv:1601.04568.
[5] Nguyen, T., Mori, K., and Thawonmas, R., 2016. Image Colorization Using a Deep Convolutional Neural Network. In Proc. of ASIAGRAPH 2016, Toyama, Japan (pp. 49-50).

Dating Mining into the Works of Monkan (1278-1357), a Monk of the Shingon School: Using Digital Humanities to Assess the Contested Authorship of Three Religious Texts

Gaetan Rappo (Waseda University)

Temples belonging to the esoteric schools of Buddhism in medieval Japan have produced an extremely vast and elaborate religious literature, called the “sacred teachings” (shôgyô). Although such texts have received little attention from historians, the last few years have seen, especially with the research team based in the Shinpukuji temple in Nagoya, a renewed interest in them as well as the publication of many manuscripts. However, the use of such texts as historical sources presents a series of challenges. The “sacred teachings” were mostly composed in Japanese kanbun (classical Chinese read in the Japanese word order), and consist in ritual procedures, doctrinal or canonical commentaries, or records of oral traditions and various events. Their writing style is fragmentary, almost cryptic at times, and they are not designed to be comprehended by the non-initiated. So, in order to understand their contents and to use them as proper historical sources, one has to reconstruct the vast knowledge network they were built upon.

Digital humanities can provide particularly helpful tools to clarify this stream of ideas and to determine patterns in the diffusion of rituals, symbols, or doctrinal interpretations among the monks of the time. They can also help us assess the veracity of traditional claims of authorship, or even evaluate the boundaries between schools or rival branches. For example, in a book published in 2015, Ishii Kôsei was able to prove the authenticity of several texts attributed to Shôtoku Taishi, relying on computer-assisted vocabulary analyses based on text mining and especially n-gram. This presentation aims to build on his research and to devise a method and tools applicable to the study of medieval “sacred teachings.” It will first examine the issue of the authorship of two texts, the Daijingû honji and the Ben’ichisan kuketsu, and assess its implications.

According to their colophons, the Ben’ichisan kuketsu and the Daijingû honji were written around the middle of the 13th century. However they both mention a ritual called the "Ritual combining the Three Worthies (Sanzon gôgyô hô)," a practice that recent scholarship has proved to have been created by Monkan (1278-1357), a monk active mostly during the 14th century. In the medieval esoteric schools, ritual and doctrinal authenticity has always been a major issue, and monks imagined various ways to assess their superiority over rival schools and practices. One of them was for a monk to attribute his own work to an eminent figure from the past. According to the research by Abe Yasurô, Monkan did this with the text called Goyuigô hiketsu, which was attributed, most certainly by Monkan himself, to the Daigoji master Jichiun, active in the early 13th century. So it is quite probable that a similar process was at work with the Daijingû honji and Ben’ichisan kuketsu.

The main hypothesis of this presentation, based on an analysis of the texts’ contents and their historical contexts, is that they may well have been written by Monkan, or at least by a member of the same school of thought, and attributed voluntarily to previous masters in order to assert their authenticity.

I will first create a database of Monkan’s known works, including the Goyuigô hiketsu (whose authorship will surely be proved in the process). The main goal is to use data mining in order to isolate linguistic pattern and analyze their frequency, mostly via python language n-gram scripts, and then to compare such data with the Ben’ichisan kuketsu and the Daijingû honji. However, occurrences of series of Chinese characters do not always have the same relevance, and, as a preliminary step, it will be necessary to organize the data.

Concretely, I will use recent editions of manuscripts and transcribe them into digital data, following the guidelines of the Text Encoding Initiative. Since I will work on a defined number of documents, I will try to create coherent metadata categories for this type of text (categories for “paragraphs” could include canonical citations, ritual procedures, opinions of their author, etc.) and apply them to the data. This will give a clearer picture of the significance of each linguistic pattern in Monkan’s

works, and help determine the topics discussed in the text, the type of references, and, most importantly in the writer's style.
The next step will be to confront this data with the Ben'ichisan kuketsu and Daijingû honji, which will first be analyzed with the same method. The results should give us a clearer picture of the degree of similitude, or maybe the differences, between them and other works by Monkan. If possible, we will also try to identify the author of another similar text, the Shinzô zuzôkan, as recent research by Uchida Keiichi has shown that the name appearing in its colophon might well refer to Monkan. However, this text is not fully available, and we would have to work with partial data.
As a whole, this research will not only help answer the question regarding the authorship of the aforementioned texts, but also shed new light on Monkan's thought as a whole, and especially his links to the medieval Shintô tradition, which for now remain uncharted territory. On a longer-term basis, this method will also prepare deeper investigations into the transmission process of rituals and ideas in the Middle Ages. Furthermore, if expanded to other texts from a similar period, it can clarify the actual diffusion of doctrinal arguments between rival schools, as Shingon and Tendai, or spiritual lineages inside the same school, allowing us not only to understand the position of a figure such a Monkan, who has the particularity to have been rejected as a heretic after his death, in the history of his school as well as in the history of ideas as a whole, but also to question the traditional divisions inherited from later periods and their influence on our understanding of the situation in medieval Japan.

Stylistic Analysis of Agatha Christie's Works: Comparing with Dorothy Sayers

Narumi Tsuchimura (Osaka University)

This study aims to clarify the stylistic characteristics in works of Agatha Christie, a female mystery writer in the UK, comparing with those of Dorothy Leigh Sayers, who was also a renowned contemporary female mystery writer in the UK from the same period.

Agatha Christie is one of the most successful female mystery writers in history, and her novels are read all over the world now. Dorothy Leigh Sayers is likewise one of the most successful female mystery writers, and had a relationship with Christie. Christie and Sayers are called the two mistresses of mystery. Sayers' mystery novels are not as popular as Christie's. However, according to Mori (1998), Sayers had better writing ability and was better at describing characters in novels than Christie, and leading female mystery writers today, including P. D. James and Ruth Rendell, say that Sayers, not Christie, is the ideal writer. On the other hand, Christie has often been recently criticised by contemporary mystery writers, saying that the characters in her novels are in a fixed form, and that her style is too plain. At this point Christie forms a great contrast to Sayers. The purpose of this study is to examine how Christie's style differs from Sayers' one by comparing their works using statistical analysis.

While quantitative researches on Christie's works, such as those of Lancashire & Hirst (2009) and Le et al. (2011), do exist, but these researches are based on simple statistics like word frequencies and Type/Token ratio. In addition, such researches do not deal with all of her works.

This study aimed to answer the following questions:

(1) Can we distinguish Christie's works from Sayers' by using statistical methods?

(2) What are the stylistic difference between the works of Christie and Sayers?

The data used in this study consists of Christie's (221 texts, 5,230,256 words) and Sayers' works (55 texts, 1,430,257 words). This study applies Random Forests for classifying the two writers' works and extracting characteristic words from each writer's works. Random Forests, which was proposed by Breiman (2001), is an ensemble learning method for classification and regression. In recent studies it has been used for classifying texts and authorship attribution. For example in Jin & Murakami (2007), Random Forests was employed for authorship identification of three different types of texts (novels, compositions and diaries), and it is shown that this method is more effective than other classifiers. In Tabata (2012), Random Forests was used to extract marker words that distinguish Charles Dickens from Wilkie Collins. Tabata reported that Random Forests overcame common problems in key word measures such as Log Likelihood or Chi-squared score, and Random Forests was proposed as the alternative method. Following Tabata, this study employed Random Forests to extract characteristic words that differentiate Christie from Sayers.

The variables used in Random Forests are the most frequent words. Random Forests is trained and validated on the 276 texts with different numbers of most frequent words ranging from 1000 to 100 in 100 word steps. The Christie texts and the Sayers texts were correctly classified into two different groups with an accuracy of 92.7%-95.3%. These texts were classified the most accurately with the top 600 words. Out of these, the 100 the most characteristic words were identified. The results of this analysis revealed that Christie's use of characteristic words contrasts with Sayers', especially between synonyms: Christie tends to use anyone, someone and until while Sayers tends to use anybody, somebody and till. Moreover, it revealed that Christie tends to use words related to the female gender, movement and visual. We discuss how these words are used differently in the texts of the two authors, and attempt to reveal the stylistic characteristics of Christie when compared with her contemporary.

References

[1] Breiman, L. (2001). Random forests. Machine Learning, 45: pp.5-23.

[2] Jin, M. and Murakami, M. (2007). Authorship Identification Using Random Forests. Proceedings of the Institute of Statistical Mathematics, 55(2): pp.255-26

[3] Lancasire, I. and Hirst, G. (2009, March). Vocabulary Changes in Agatha Christie's Mysteries as an Indication of Dementia: A Case Study. Paper presented at the 19th Annual Rotman Research Institute Conference, Cognitive Aging: Research and Practice, Toronto.
[4] Le, X., Lancashire, I., Hirst, G. and Jokel, R. (2011). Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists. Literary and Linguistic Computing, 26(4): pp.435-461.
[5] Mori, H. (1998). Sekai Mystery Sakka Jiten: Honkakuha-hen. Tokyo: Tosho Kankoukai.
[6] Tabata, T. (2012). Stylometry of co-authorship: Charles Dickens and Wilkie Collins. The Special Interest Group Technical Reports of Information Processing Society of Japan, CH-93(3): pp.1-7.

Jane Austen in Vector Space: Applying vector space models to 19th century literature.

Sara J Kerr (Maynooth University)

Jane Austen is known as one of the founders of the modern English novel. Traditionally she has been seen as a writer who focused on the minutiae of domestic life, but more recent critics have been finding ideas which challenge this view, positioning her as a far more political writer than previously thought.

An exploration of Austen's ideology is challenging for a number of reasons. A lack of contextual material means that no reliable, first person account exists. The historical representation of Austen has been heavily mediated by her family. With the publication of A Memoir of Jane Austen in 1869, some fifty years after her death by her nephew James Edward Austen-Leigh, Austen's reputation as a skilled but uncontroversial writer, in keeping with the Victorian ideal of conservative, religious, womanhood, was set.

Yet, there is a disconnect between the Austen presented by her family and the Austen we glimpse through her novels. Most modern scholars agree that there are some political elements in Austen's work, but there is considerable disagreement as to their nature. Austen was writing with the goal of publication and was clearly aware that publishers and the public had clear expectations of novels and that openly contentious works were unlikely to be published. Her writing, first and foremost, aimed to entertain and as a result, it is perhaps unsurprising that her political ideology is to hard to identify. This research takes a quantitative view of Austen's novels, focusing on her representation of independence and dependence at a personal, social and political level, to explore her political ideas in detail.

Traditional close reading by necessity focuses on the detailed analysis of small sections of text. Although the corpus of Austen's novels is not large, scholars have identified very different political views using the same source material. For example, Butler (Jane Austen and the War of Ideas. Oxford: Clarendon Press, 1997. Print.) presents Austen as a conservative writer whereas Johnson (Jane Austen: Women, Politics, and the Novel. University of Chicago Press, 1990. Print.) and Neill (The Politics of Jane Austen. Macmillan, 1999. Print.) claim that she is more radical in her views. In searching for insight into the traces of Austen's political views, we need to look for more subtle patterns within and across the texts. In effect, we are looking for an understanding which goes beyond the individual novel, and in Moretti's words "close reading will not do it" (Distant Reading. London: Verso, 2013. Print. p48).

The advent of distant and scaled reading techniques within literary studies has enabled the exploration of texts in a manner which "defamiliarize…making them unrecognizable in a way…that helps scholars identify features they might not otherwise have seen" (Clement, Tanya. "Text Analysis, Data Mining and Visualisations in Literary Scholarship." MLA Commons | Literary studies in the digital age. Oct. 2013. Web.). Topic modelling is, perhaps, the most popular of these tools for Digital Humanists who wish to transform texts and view them through a different lens. However, the application of 'word2vec' (an algorithm which represents words as points in space, and the meanings and relationships between them as vectors) has the potential to be of even greater use. It can work effectively on a smaller corpus and can be applied to full texts, whereas, as Jockers has noted ("'Secret' recipe for topic modeling themes'. matthewjockers.net. 12 Apr. 2013, Web.), topic modelling is more effective when working with a large, noun only corpus. In addition, 'word2vec' allows the exploration of discourses surrounding a theme. Rather than asking 'which topics or themes are in this corpus of texts?' the application of the 'word2vec' algorithm allows us to ask 'what does the corpus say about this theme?'

Typical results are illustrated through an investigation of the words 'independent' and 'independence'. The ten nearest words to 'independent' include 'decorum', 'fortune' and 'greatness', but also 'contemptible' and 'deplorable'. Similarly, the fifty nearest words to 'independent' and 'independence' contain many of the expected words: 'privilege', 'matrimonial', 'property', 'fortune', however, also present a different group: 'inequality', 'littleness', 'illiberal' and 'unfair'. This group of words suggests that while much of Austen's discourse surrounding independence is in keeping with views of her writing as conservative, a more critical discourse also exists.

Expanding the number of words in the model to one hundred reveals two of these negative clusters:

1. bias, inequality, insignificance, wounding, littleness, unfair, degradation, shameful
2. deplorable, arrogance, contemptible, conceit, spoiled

This is suggestive of two separate ideas being expressed: the negative impact of the unfair distribution of wealth on the less fortunate, and the negative impact independence could have on its recipient. As wealth and property underpinned the existing social hierarchy these views may be seen as political.

The 'word2vec' model, originally created by Tomas Mikolov and his colleagues at Google in 2013, takes in a corpus of texts and represents words as points in a multi-dimensional space, word meanings and relationships between words are encoded as distances and paths in that space, through the creation of an artificial neural network. The created model can then be interrogated and the results visualised.

Applying 'word2vec' to literary studies allows the discursive space surrounding a particular topic to be examined, highlighting areas for further exploration. While close readings can identify specific examples where Austen is critical of the world in which she lives, the application of 'word2vec' suggests that a more consistent discourse critical to the existing power structures exists across her novels. An exploration of the novel corpus within vector space places the discourse within a semantic space through which Austen's ideological views can be interrogated in combination with a more traditional close reading, leading to a thicker, more nuanced interpretation.

Although a relatively recent addition to the range of computational tools being used for Digital Humanities, these initial analyses suggest that there is good reason to explore the application of vector space models to corpora of literary texts further.

MEDEA (Modeling semantically Enhanced Digital Edition of Accounts) as Historical Method

Kathryn Tomasek (Wheaton College)

Introduction

A cooperative project among historians in Germany, Austria, and the United States who are interested in developing models for digital scholarly edition of account books for comparative historical analysis, MEDEA was formed in 2014. Our goals are data modeling and expanding the community of practice for these activities. We have spent the past year introducing these ideas to scholars in Europe and the United States, holding workshops in Regensburg in October 2015 and at Wheaton College in Massachusetts in April 2016. Georg Vogeler Professor of Digital Humanities at the Austrian Centre for Digital Humanities, Centre for Information Modeling at Karl Franzens University in Graz is testing files created by other members of our community against his bookkeeping ontology.

Problem Space: Accounts

Accounts of various sorts—municipal, state, organizational, merchant, and individual—are abundant in archives, but they are underutilized as sources at least in part because of the technologies that have been used to produce and analyze them. They have been important sources both for those who created them and for historians who have sought to understand economic changes over time and space. Historians have sampled such sources using social science methodologies for almost one hundred years, but few scholarly editions of accounts have been produced (Ciula, Spence & Veira 2008, Keating et al. 2010, Teehan & Keating 2010, Bolt & van Zanden 2014, Frantz & Sarnowsky 2014, Burghartz 2015).

At least some of the challenges in producing digital scholarly editions of accounts are related to the development of the very technologies that have been used to record accounts in the past several hundred years in the global North. Account books—codices containing lists of commodities, currencies, and services exchanged among people—developed over time into printed ledgers and spreadsheets—analog books and papers that could be used to record information about these exchanges in tabular format. The formats of the various cross-referenced books of accounts associated with the business of running cities, estates, mercantile operations, and other enterprises gave people opportunities to track inventories, obligations, and assets with a view to such questions as personal, organizational, or state or municipal wealth. (Discussion of accounts kept on clay tablets, papyri, scrolls, and other media that preceded the codex are omitted only as a reflection that MEDEA currently does not include any scholars who are working with sources in such forms; we welcome such scholars and the challenges their sources will bring.) Accounting practices are themselves a technology that have undergone changes over time and space (Ijiri 1975, Everest & Weber 1977, Bywater & Yamey 1982, McCarthy 1982, Wigley n.d., Mersiowsky 2000, Wang, Du & Lee 2002, Arlinghaus 2004, Vogeler 2005, Vogeler 2010).

Digital versions of this analog technology in the form of spreadsheet software, relational databases, and web-based forms, such as the business software XBRL-General Ledger, have the advantage of simplifying the tracking of sums, balances, and in fact most numerical or mathematical operations as well as producing visualizations. However, spreadsheet software handles semantic values much less efficiently. Information about which currencies, commodities, services, individuals, and geographical locations are referenced in exchanges between groups or individuals can easily be lost or misrepresented in spreadsheets. Even more flexible relational databases are often idiosyncratic in their references to such semantic values and fail to meet any sorts of standards for interoperability, despite the considerable social scientific literature based on sampling from analog sources. Oxford historian Richard C. Allen has undertaken to assess the quality of data extracted from accounts and electronically available (Allen 2001, Allen et al. 2004, Allen 2014).

Possible Solutions: An Event-Based Ontology for Accounts

Vogeler and Tomasek have been working on somewhat parallel paths for the past decade or so. Both began from the position of the Guidelines of the Text Encoding Initiative (TEI) as an accepted method for producing stable humanities-oriented data, and both have sought to leverage the TEI's position as a standard for such work to explore models for creating reusable and interoperative digital scholarly editions of accounts from original sources distant in time and space. (Tomasek & Bauman 2013, Vogeler 2015).

Vogeler has outlined a preliminary version of an RDF model for comparing accounts. In describing this model, he has argued that the "transactionography" TEI customization that Tomasek and Bauman developed a few years ago amounted "a simple ontology for accounting facts" (Vogeler 2016). Thus, the ontology that he has been developing begins from the transaction, incorporating the notion that

a transaction between two parties or accounts consists of at least one transfer from one to the other. It transfers a measurable and can be attested by text. The transfer occurs at a place. Booking a transfer into an account can create liabilities held by a party and owed to another (Vogeler 2016).

Vogeler borrows additional data types from XBRL-GL and TEI. These include monetary values, the entry, debit/credit, the balance, totals, and measure. And attending to the interests of historians, he adds prices, commodities, services, and conversions of measurements. Vogeler suggests further that common terms from the taxonomies developed by individual projects can be identified, exposed as RDF data, and described using the W3C's Simple Knowledge Organization System (SKOS).

Along with Øyvind Eide, Vogeler presented slides at DH2016 that draw on CIDIC-CRM's event-based modeling to point towards an ontology that can express both the human interactions and the accounting practices represented in account books (Eide & Orr 2009). Vogeler's slide outlines the production of accounts as traces of human activity, historians' interests in what accounts can tell us about the past, and the technologies most appropriate to creating digital surrogates susceptible to analysis. Eide's slide illustrates how an event-based model of the activities that produced accounts can be expressed using principles from CIDOC-CRM. And Vogeler's sketch of his bookkeeping ontology on GitHub offers a picture of its current status.

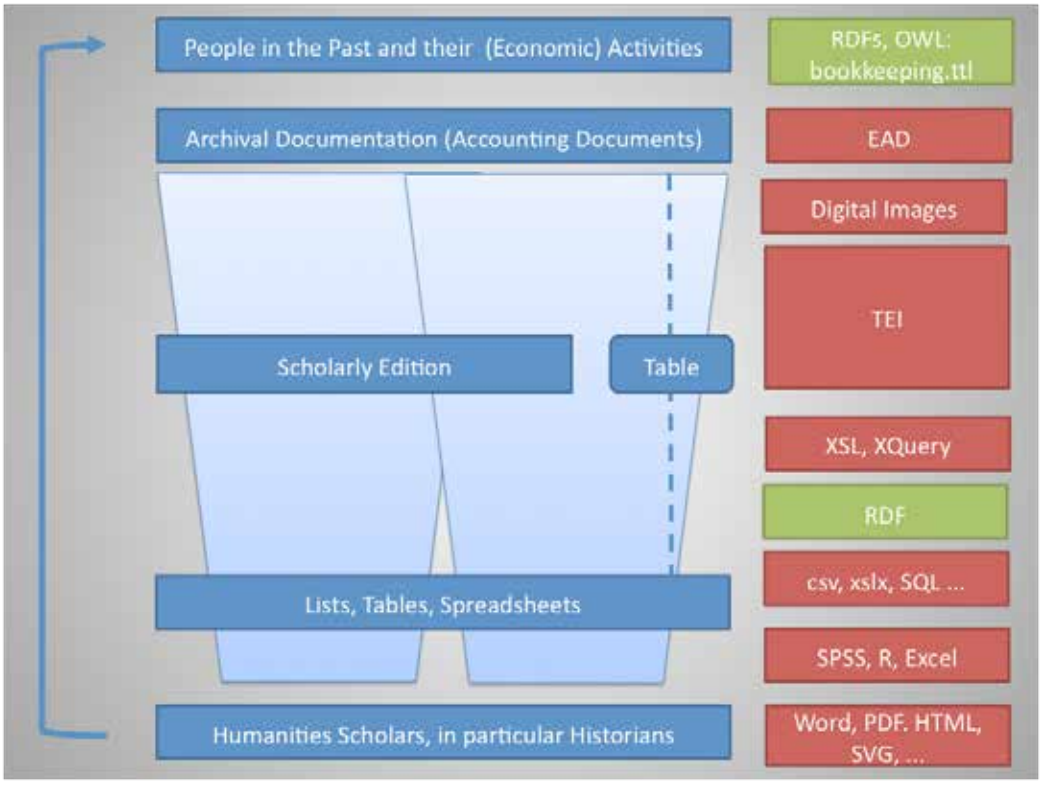

Fig 1. (Image Credit: Georg Vogeler, DH2016.)

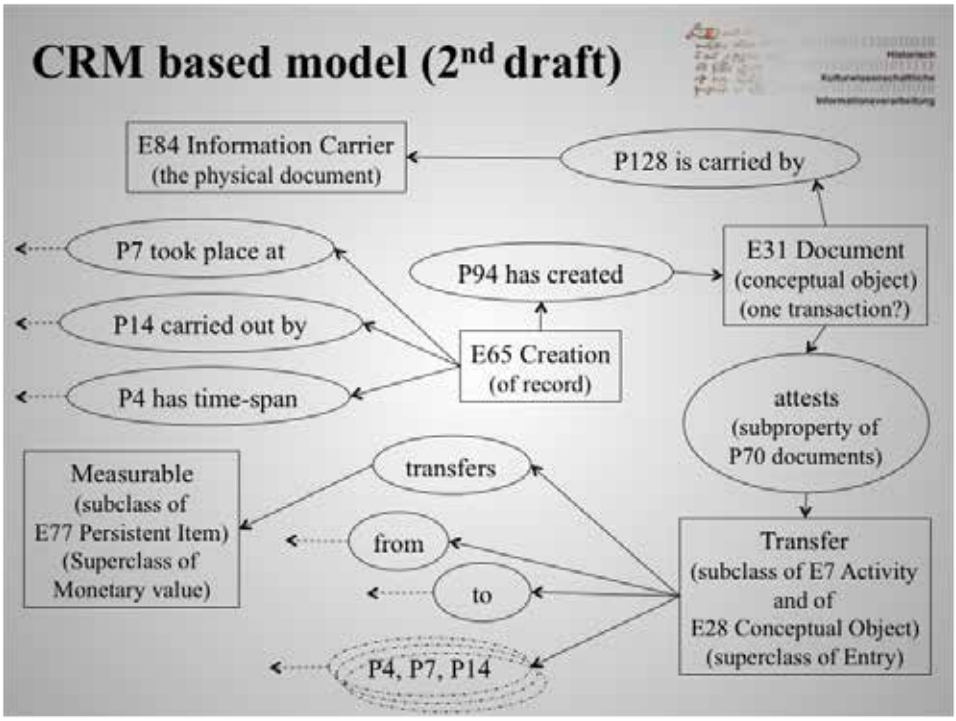

Fig 2. (Image Credit: Øyvind Eide, DH 2016.)

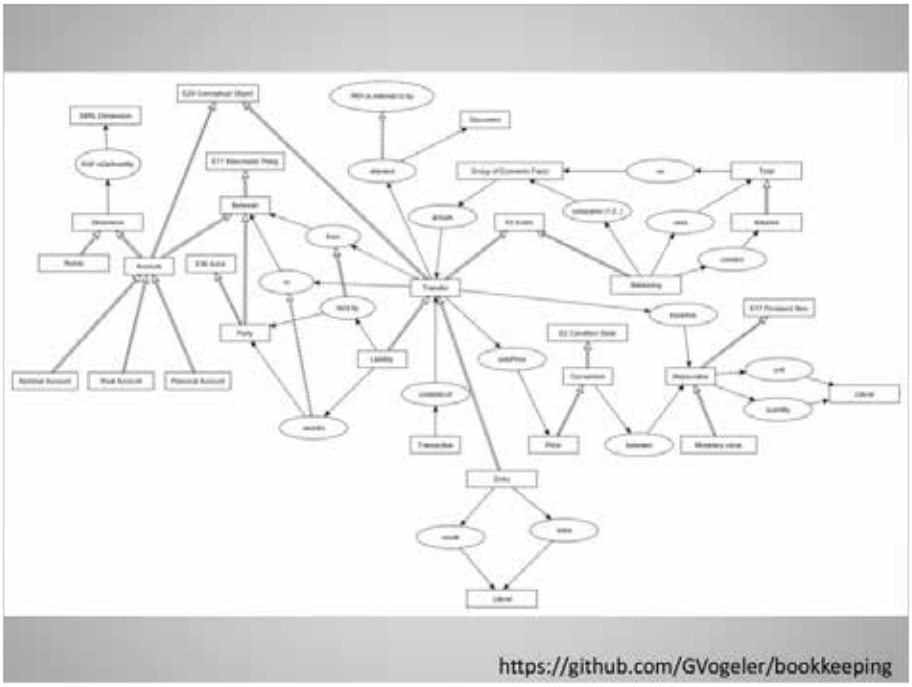

Fig 3. (Image Credit: Georg Vogeler, DH 2016.)

Currently, Vogeler suggests using the TEI @ana attribute to add markup for this bookkeeping ontology. Such markup bypasses the need for the kind of "transactionography" described by Tomasek and Bauman, allowing markup of such information as "transfer," "from," "to," or the ambiguous "between," as well as "monetary value," "what" was transferred, and whether the transfer was "mutual," "multiple," "unilateral," or "enforced." Example markup from my own project will show this ontology in use.

Following best practices for digital scholarly editions, the XML/TEI file can be stored, either alongside images of the original archival documents or with pointers to them. The bookkeeping information from the XML/TEI can be converted to RDF for comparison to other documents

marked up in similar manner. Vogeler has tested such comparisons with a small set of files and is eager to increase the number of files marked up in such manner for further testing (Vogeler 2016).

Conclusions

Widespread scholarly edition of accounts using the ontology that Vogeler and Eide are developing has the potential eventually to offer an unprecedented source of aggregated data for historical research. As a result of MEDEA workshops, we have added to the community of practice for transcription and markup of accounts following Vogeler's recommendations for use of XML/TEI with RDF using his bookkeeping ontology. Along with some colleagues in the United States, I am currently seeking funding to encourage use of Vogeler's rmodel both in educational contexts and in those occupied by citizen archivists. We hope thus to increase the number of accounts available for comparison and to demonstrate the advantages of digital scholarly edition of accounts for making available historical data that can be reused by other scholars. Described in this way, our goals nothing new in Digital Humanities with regard to digital scholarly edition of texts. In its focus on accounts however, MEDEA marks a significant new opportunity. MEDEA challenges historians especially to consider digital scholarly edition of accounts as a new model of scholarship that takes advantage of the affordances of the Semantic Web.

Note

A portion of the activities described in this paper is supported jointly by the National Endowment for the Humanities and the Deutsche Forschungsgemeinschaft. Any views, findings, conclusions, or recommendations expressed in this paper do not necessarily reflect those of the National Endowment for the Humanities or the Deutsche Forschungsgemeinschaft.

References

[1] Allen, R. C. 2001: The Great Divergence in European Wages and Prices from the Middle Ages to the First World War. In: Explorations in Economic History 38, S. 411–447.

[2] Allen, R.C. 2013. The High wage Economy and the Industrial Revolution. A Restatement. In: Oxford Economic and Social History Working Papers (Ref: Number 115), http://www.economics.ox.ac.uk/index.php/Oxford-Economic-and-Social-History-Working-Papers/the-high-wage-economy-and-the-industrial-revolution-a-restatement.

[3] Allen, R.C., Clark, G., Devereux, J., Hellie, R., Hoffman, P.T., Jacks, D. S., Lindert, P. H., Ma, D., Mironov, B. N., Pamuk, S., Van Zanden, J. L., Ward, M. 2004. Preliminary Global Price Comparisons 1500-1870. In: Lindert, P. H. et. al.: Towards a Global History of Prices and Wages, 19-21 Aug. 2004, <http://www.iisg.nl/hpw/conference.html>

[4] Arlinghaus, Franz-Josef. 2004. Bookkeeping, Double-Entry Bookkeeping, in: Medieval Italy. An Encyclopedia. Ed. Christopher Kleinhenz, John W.Barker, Gail Geiger, Richard Lansing. Routledge, 01-08-2004. Routledge History Online. Taylor & Francis. Accessed 19 July 2016 <http://routledgeonline.com:80/history/Book.aspx?id=w440>

[5] Bolt, J. and J. L. van Zanden (2014). The Maddison Project: collaborative research on historical national accounts. The Economic History Review, 67 (3): 627–651.

[6] Burghartz, Susanna, ed. 2015. Jahrrechnungen der Stadt Basel digital, unter Mitarbeit von Sonia Calvi, Lukas Meili, Jonas Sagelsdorffer und Georg Vogeler, Basel/Graz: Zentrum für Informationsmodellierung der Universität Graz.

[7] Bywater, M.F. and B.S. Yamey. 1982. Historic Accounting Literature: A Companion Guide. London: Scholar Press.

[8] Ciula, Ariana, Paul Spence, and José Miguel Veira. 2008. Expressing complex associations in medieval historical documents. The Henry III Fine Rolls Project, in: LLC 23:311-325.

[9] Eide, Øyvind and Christian-Emil Ore. 2009. "TEI and cultural heritage ontologies." LLC 24,2:161-172.

[10] Everest, Gordon C., and Ron Weber. 1977. "A Relational Approach to Accounting Models." The Accounting Review 52, no. 2: 340–59.

[11] Franzke, C. u. J. Sarnowsky. 2015. Amtsbücher des Deutschen Ordens um 1450. Pflegeamt zu Seehesten und Vogtei zu Leipe. Göttingen: V & R unipress, (Beihefte zum Preußischen Urkundenbuch 3).

[12] Graham, Shawn, Ian Milligan, and Scott Weingart. 2016. Exploring Big Historical Data: The Historian's Macroscope. London: Imperial College Press.
[13] Ijiri, Yuji. 1975. Theory of Accounting Measurement. Studies in Accounting Research 10. Sarasota, FL: American Accounting Association.
[14] Jones, Eric. 1981. The European Miracle: Environments, Economies, and Geopolitics in the History of Europe and Asia. Cambridge: Cambridge University Press.
[15] Keating, John G., Aja Teehan, Damien Callagher, and Thomas O'Connor. 2010. A Digital Edition of a Spanish 18th Century Account Book. User Driven Digitisation, in: Computerphilologie 10:169-188.
[16] McCarthy, William E. 1982. "The REA Accounting Model: A Generalized Framework for Accounting Systems in a Shared Data Environment." The Accounting Review 57, no. 3: 554–78.
[17] McCusker, John J. 2001. How Much Is That In Real Money?: A Historical Commodity Price Index for Use as a Deflator of Money Values in the Economy of the United States, 2d ed., rev. and enlarged. Worcester, Massachusetts: American Antiquarian Society.
[18] Mersiowsky, Mark. 2000. Die Anfänge territorialer Rechnungslegung im deutschen Nordwesten. Spätmittelalterliche Rechnungen, Verwaltungspraxis, Hof und Territorium (zugl. Diss. phil. Münster 1992) , Sigmaringen: Thorbecke, 2000 (Residenzenforschung 9).
[19] Polanyi, Karl. 1944. The Great Transformation. New York: Farrar and Rinehart. Pribram, A. F. (1938): Materialien zur Geschichte der Preise und Löhne in Österreich. Band 1. Wien: Carl Ueberreuter Verlag.
[20] Poovey, Mary. 2008. Genres of the Credit Economy: Mediating Value in Eighteenth- and Nineteenth-Century Britain. Chicago, Ill.: University of Chicago Press.
[21] Teehan, Aja and John G. Keating. 2010. A Digital Edition of a Spanish 18th Century Account Book. Part 2 - Formalisation and Encoding, in: Computerphilologie 10:189-214.
[22] Tomasek, Kathryn 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 10 April 2016. URL : http://jtei.revues.org/895 ; DOI : 10.4000/jtei.895
[23] Vogeler, Georg. 2005. Tax Accounting in the Late Medieval German Territorial States, in: Accounting, Business and Financial History 15 , S. 235-254.
[24] Vogeler, Georg. 2010. Financial and Tax Reports, in: de Gruyter Handbook of Medieval Studies. Concepts, Methods, Historical Developments, and Current Trends in Medieval Studies, hg. v. Albrecht Classen, Berlin, S. 1775-1784
[25] Vogeler, Georg. 2015. Warum werden mittelalterliche und frühneuzeitliche Rechnungsbücher eigentlich nicht digital ediert?, in: Grenzen und Möglichkeiten der Digital Humanities, hg. v. Constanze Baum u. Thomas Stäcker , Wolfenbüttel. DOI: 10.17175/sb001_007, URL: <http://zfdg.de/warum-werden-mittelalterliche-und-fr%C3%BChneuzeitliche-rechnungsb%C3%BCcher-eigentlich-nicht-digital-ediert>.
[26]Vogeler, Georg. 2016. The Content of Accounts and Registers in their Digital Edition. XML/TEI, Spreadsheets, and Semantic Web Technologies, in: Edition von Rechnungen und Amtsbüchern, hg. v. Jürgen Sarnowksy, S. im Druck.
[27]Wang, Ting J., Hui Du, and Hur-Li Lee. 2002. "A User-Oriented Approach to Data Modeling: A Blueprint for Generating Financial Statements and Other Accounting-Related Documents and Reports." The Review of Business Information Systems 6(4): 17–32. DOI:10.19030/rbis.v6i4.4548
[28] Wigley, Michael. n.d. Double Entry Accounting in a Relational Database. http://homepages.tcp.co.uk/~m-wigley/gc_wp_ded.html. Accessed August 15, 2013.

Modeling New TEI/XML Attributes for the Semantic Markup of Historical Transactions, based on 'Transactionography'

Naoki Kokaze (University of Tokyo), Kiyonori Nagasaki (International Institute for Digital Humanities), Masahiro Shimoda, A. Charles Muller (University of Tokyo)

This presentation addresses the provision of extended TEI/XML attributes based on 'Transactionography' to mark up the historical transactions more semantically and precisely.
'Transactionography' is a methodological approach proposed by Drs. Kathryn Tomasek and Syd Bauman in their article published in 2013, 'Encoding Financial Records for Historical Research.' This methodology is an extended model of the TEI in order to mark up appropriately various kinds of historical financial records containing histories of the exchange of goods and services, such as receipts or account ledgers.
They focused on micro transactions relating to personal life in the article. But, naturally enough, a transaction is included in trade which is also performed between states as well as in the individual realm. When marking up those macro transactions, we expect that the issues of how to tag the states, the individuals, which were the subjects of the transactions, or the transactions themselves, remains as problems.
To solve these issues, at the Second MEDEA workshop held in April 2016 at the Wheaton College, we already presented a concrete example of how to utilize the TEI attributes, especially @sameAs and @corresp, with the new defined elements of 'Transactionography'.
In that presentation, we took the transactions of gunboats, the so-called Lay-Osborn Flotilla, between the governments of Great Britain and China in the 1860s, as an example. The purpose of those transactions for the Chinese government was to, with minimal financial cost, to suppress the pirates and rebels in Chinese territory without building a naval force, by introducing the advanced Western military technology. Whereas the purpose for the British government was to decrease the burden of the duty required to maintain the order of the China seas instead of the Chinese maritime force, and also to reduce naval expenditures. However, the flotilla was eventually disbanded, because of a difference in their opinions about whether making use of the Lay-Osborn Flotilla for the purpose of strengthening the Chinese central power or Chinese local power.
In these transactions, a key person was the Inspector General, Lay. Having returned temporarily to England for medical treatment, Lay, rather than the Chinese Government, purchased the fleets and weapons there. In addition to these expenses, fleet maintenance costs and the salary of the sailors on board, which were estimated to be necessary after their arrival in China, were paid from each Chinese Maritime Custom as bills, being sent to Lay from time to time. However, as already mentioned, since the Lay-Osborn Flotilla ended up being disbanded without any achievement, the Chinese government furiously instructed Lay to reimburse all the money he had at the moment to the government and to submit a detailed account of funds he had used so far.
When attempting to grasp the overall picture of such complex transactions, one of the merits of the application of 'Transactionography' is to make it easier to visualize a structure of transactions by giving an xml:id to each transaction or to a series of transactions that have a certain point in common. The @fra and @til attributes are very powerful in terms of connecting the actors of transactions.
But, taking into consideration international trade, there is a case where it is necessary to subdivide the trade into trading of smaller scale. In this case of the Lay-Osborn Flotilla, the transactions can be listed as follows:
(a) Purchase of fleets and ammunitions: Lay ⇆ Great Britain
(b) Remittance for the above transaction: Chinese Maritime Customs → Lay & Hart, the Acting Inspector General during the absent of Lay
(c) Reimbursement after the dissolution of the flotilla: Lay → Chinese Government

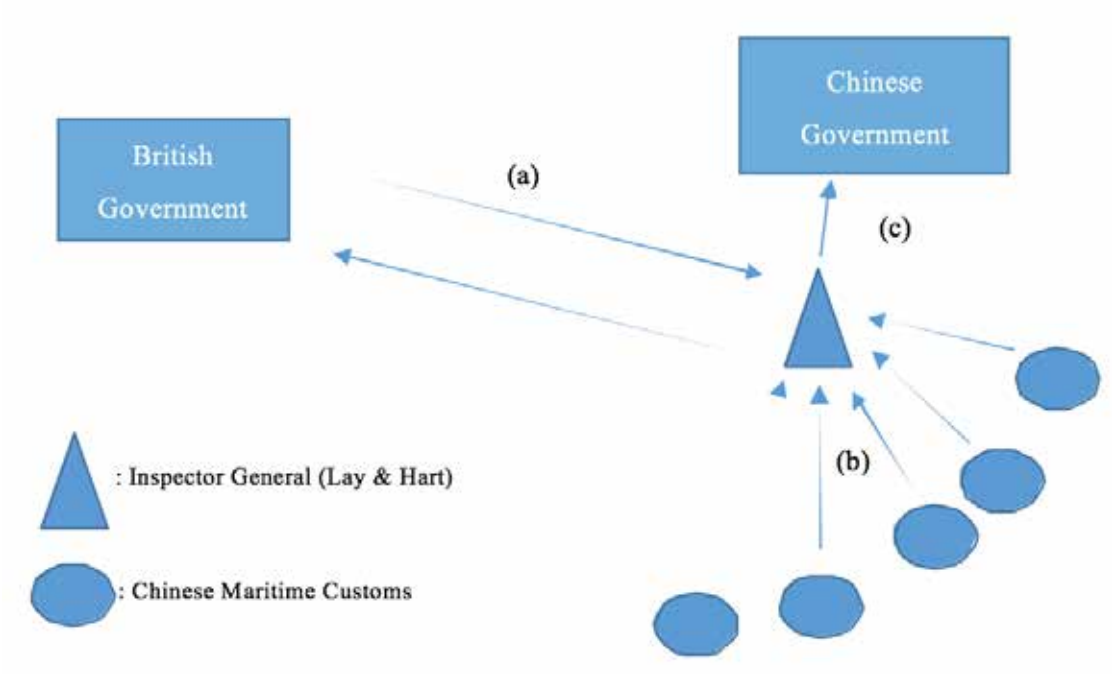

Fig 1.

What is particularly important is that Lay sent all the money he had received from each Maritime Custom directly to the Chinese government. In this chart, the cluster (b) of transactions needs to be associated with the cluster (c). Therefore, we have marked up as follows.

```
<hfrs:listTransaction xml:id="tr002"><!- Remittance to Lay -->
      <hfrs:transaction>
            <hfrs:transfer fra="#gb-hart" til="#gb-lay">
                  <measure sameAs="#ch-pay006"/>
              <!- <measure commodity="bill" quantity="20000"
                              unit="tael" xml:id="ch-
pay006">20,000 taels</measure> -->
            </hfrs:transfer>
      </hfrs:transaction>
      <hfrs:transaction>
            <hfrs:transfer fra="#ch-amoy" til="#gb-lay">
                  <measure sameAs="#ch-pay007"/>
            </hfrs:transfer>
      </hfrs:transaction>
</hfrs:listTransaction>
```

Fig 2.

As you can see here, gathering together a chain of remittance from Maritime Customs to Lay, we defined them as a single hfrs:listTransaction with an xml:id="tr002". In the Markup, we have already defined the information about the measure element with xml:id, such as "#ch-pay006", in the text above this hfrs:listTransaction element, so as not to commit a syntax error. In other words, we have separated the information about the transactions from the Markup of prose text itself.
Furthermore, as listed underneath, on the basis of this "#tr002" hfrs:listTransaction, we define a new hfrs:listTransaction with an xml:id="tr003" to clarify the relationship between those clusters of transactions.

```
<hfrs:listTransaction xml:id="tr003" corresp="#tr002">
        <!– Lay to the Chinese Government -->
        <hfrs:transaction>
            <hfrs:transfer fra="#gb-lay" til="#ch-gov">
                <measure corresp="#lay-pay001"/>
            </hfrs:transfer>
        </hfrs:transaction>
        <hfrs:transaction>
            <hfrs:transfer fra="#gb-lay" til="#ch-gov">
                <measure corresp="#lay-pay002"/>
            </hfrs:transfer>
        </hfrs:transaction>
</hfrs:listTransaction>
```

Fig 3.

Thus, it is helpful to use @corresp attribute in connecting related transactions with their xml:id. In this way, we should be able to describe the whole structure of transactions, even though it is complicated, by breaking down international trade into exchanges of smaller scale.
In addition, we have constructed an interface for the visualization of the flow of money and for the multi-lingual historical approach, which is going to be published in the papers of IPSJ SIG notes in 2016.

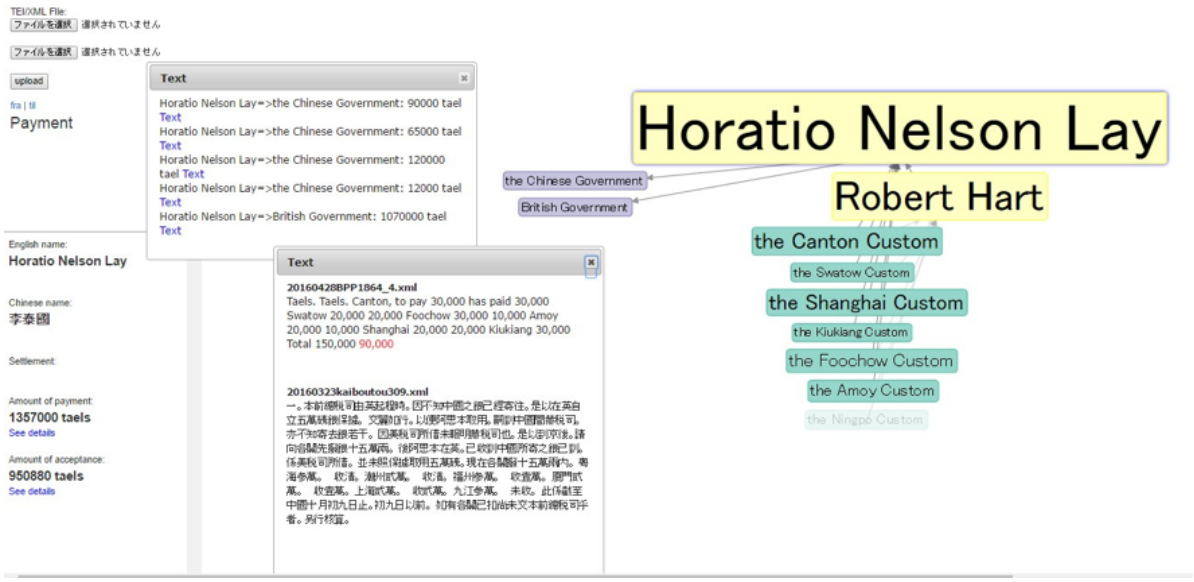

Fig 4.

However, there still remains a further problem that the @corresp attribute is too ambiguous to describe the relationship between each of the transactions. For the purpose of the semantic markup, e. g., when the total sum of a series of transactions would be the basis for other transactions, it should be helpful to define new attributes, such as @sum. In this coming presentation at JADH 2016, we would like to suggest new attributes to mark up historical transactions informatively and semantically, by collecting a variety of samples.

HYU:MA -– A Model for Library-–Supported Projects in Japanese Digital History

Peter Broadwell, Tomoko Bialock (University of California, Los Angeles)

This presentation provides an overview of a cluster of new digital research projects at UCLA that focus upon Japanese historical literature, visual arts, and theater traditions. These projects are noteworthy for their use of emerging technologies in digital scholarship and instruction, as well as their basis in an active partnership between faculty members and the university library — specifically, between subject librarians, library technology specialists, and professors in the department of Asian Languages and Cultures. In addition to contributing new digital scholarship tools and resources, research findings, and instructional methods, these projects may demonstrate a promising alternative model for enabling more sustainable and collaborative digital history projects both within and across institutions.

The projects described here are meant to develop novel ways of making Japanese cultural history more accessible and compelling as fields of study for new generations of undergraduates and graduate students. They use as a guiding metaphor the emergent notion of the digital "humanities macroscope" (stylized as ヒューマ, or HYU:MA) — an infrastructure for research and instruction that facilitates computational access to large historical corpora, tools and visualizations. In doing so, the macroscope enables exploration of historical phenomena across a range of perspectives: from close reading (micro-scale), to distant reading (macro-scale), and at all levels in between (meso-scale).

The forerunner of the new generation of digital projects in Japanese cultural history at UCLA is the Hentaigana App, a highly successful collaborative effort between faculty, library staff and students at UCLA and Waseda University to develop an app for iOS and Android smartphones that helps students learn to read premodern Japanese calligraphic writing.[1] The highly interactive, customizable and even entertaining features of the Hentaigana App allow scholars very quickly to develop substantial facility for reading historical manuscripts. Experiencing Japanese classical literature in this form, rather than solely from typeset critical editions, enables scholars to engage much more closely with the original historical context of the texts and encourages them to ask more detailed questions about the circumstances that produced them. Importantly, the app uses actual historical texts selected and digitally curated by librarians and faculty at Waseda and UCLA as the source materials for its interactive lessons.

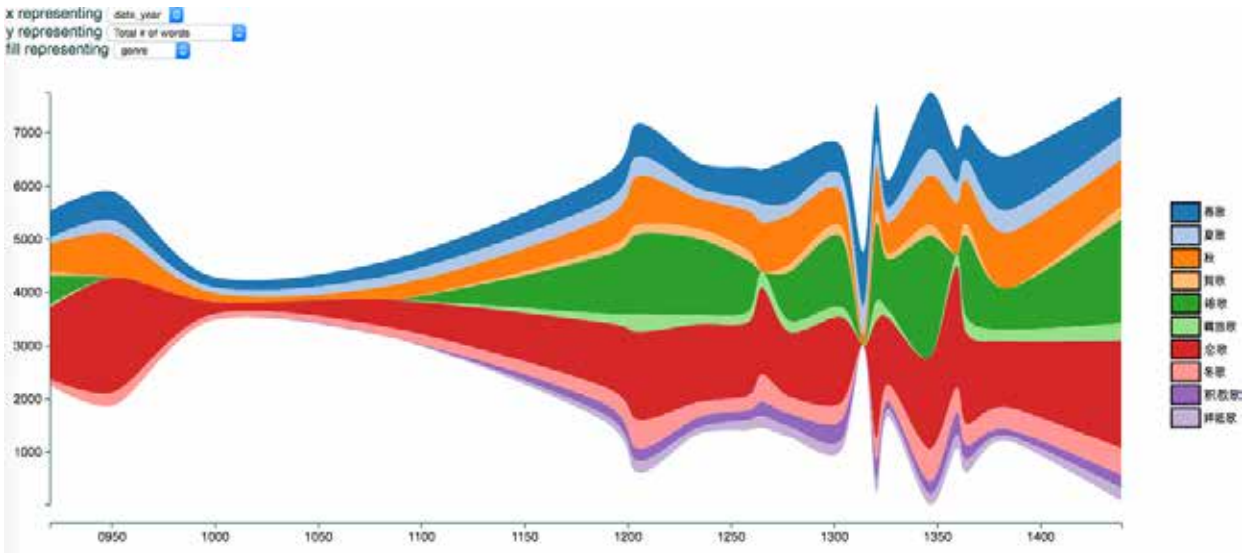

Fig 1. An advanced n-gram search and visualization interface for Japanese poetic texts, based on the Bookworm open-source project (http://bookworm.culturomics.org/).

The HYU:MA tools developed subsequent to the Hentaigana App follow the model it exemplifies: employing modern digital technologies to provide new perspectives on historical Japanese cultural expressive forms — primarily poetry and prose fiction, but also aspects of visual culture — while

making use of digitized library collections and benefitting from library staff members' increased proficiency with the development and use of digital scholarship tools. Among such products is an interactive resource, based on the open-source Bookworm "n-gram" viewer project, that enables scholars to produce and browse interactive visualizations of word frequencies among tens of thousands of Japanese poems over several centuries beginning from approximately 600 AD, including the imperial anthologies of waka poems (see Figure 1).

Besides allowing scholars to gain a "distant reading" perspective on fluctuations in word usages over time via interactive kanji- and kana-based search features, the n-gram search interface provides visualizations of the contributions of individual poets and the compilations of specific anthologies, as well as the genre categories assigned to the poems within the anthologies. This latter feature enables contemporary scholars to re-engage with prior analyses of poem types and vocabularies from the early years of mixed quantitative and qualitative waka studies,[2] as well as more recent inquiries in the fields of computational corpus linguistics.[3] A related type of analysis that is presently underway involves a comparison of human-labeled poem genres to those assigned by a computational classifier that has been "trained" by observing the words most commonly associated with traditional genres. The so-called "confusion matrix" that results from this comparison highlights the poems on which human and computerized classifiers disagree (see Figure 3). These "disagreements" tend to highlight poems that do not fit neatly into a single category; in studies of other bodies of literature, such liminal cases have proven upon further, close reading to be some of the most culturally and historically significant works within the entire corpus.[4]

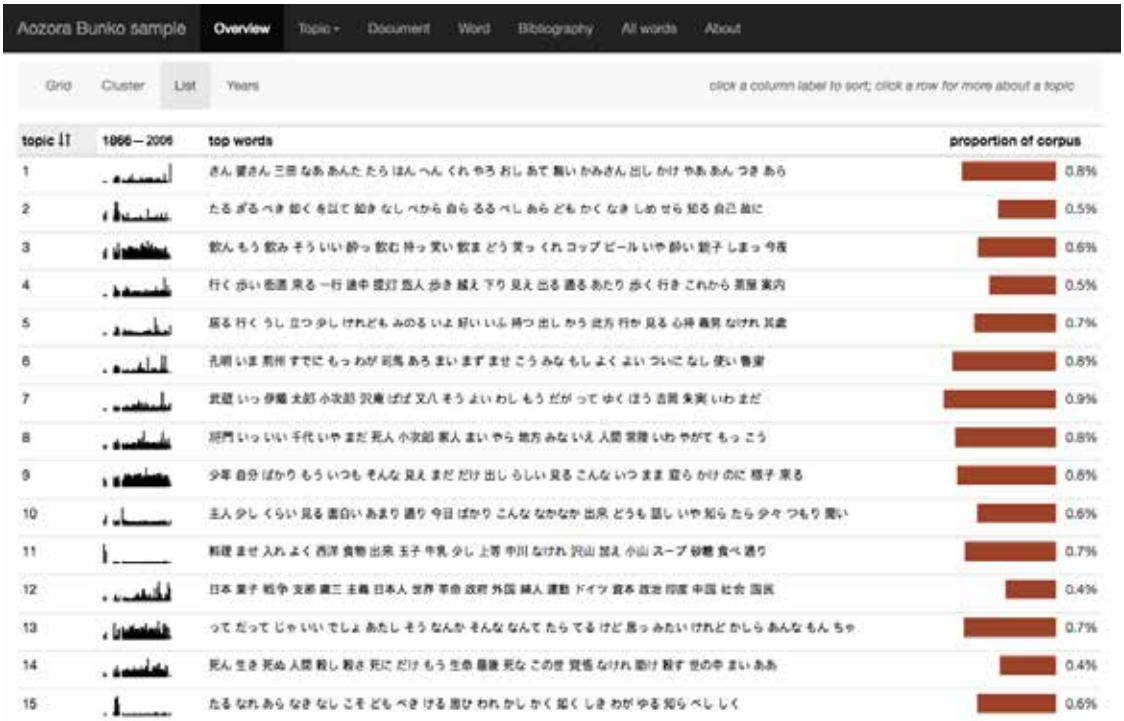

Fig 2. A sample topic model browser for 1,564 novels from Aozora Bunko, based on the open-source dfr-browser project (http://agoldst.github.io/dfr-browser/).

Further tools in development at UCLA involve the use of Latent Dirichlet Allocation-based topic modeling to generate interactive visualizations of the concentrations of various computationally detected semantic "topics" within single works (for example, Genji monogatari), as well as across very large corpora, including the publicly accessible prose materials in the Aozora Bunko digital library (see Figure 2). When applied to single works, topic modeling can reveal interesting mid-level (meso-scale) attributes of the work that may have escaped researchers' notice; when used on much larger collections that no single reader can expect realistically to read and comprehend in full, this tool provides a helpful "distant," aggregate view of the primary topics in the entire corpus and their attributes over time; scholars may then choose to examine a few specific works, topics, or time periods in greater detail.

Fig 3. A "confusion matrix" visualization of official genre classifications (vertical axis) of waka poems from the imperial anthologies, versus the classifications made by a naïve Bayes classifier trained on the official classifications (horizontal axis).

One other experimental approach that we are pursuing involves utilizing recent advances in computational image analysis to provide distant and meso-scale perspectives on large collections of images. These methods include automatic generation of image mosaics, analyses and visualizations of the color profiles of images, and even the ability to search for similar objects across multiple images. Such techniques are particularly applicable to the rich visual characteristics — which, ironically, often make text-based analytical approaches considerably more difficult — of many Japanese cultural products, from Ehon banzuke playbills to ukiyo-e prints (See Figure 4).

Fig 4. An aggregate visualization of the front pages of the UCLA Library's collection of digitized Ezukushi banzuke playbills, sorted horizontally by hue and vertically by brightness, using the open-source Coverspace tool (http://benschmidt.org/projects/coverspace/).

As a concluding observation, it is important to emphasize again the substantial and ongoing involvement of librarians and library technical staff in the digital scholarship endeavors described above. This arrangement may demonstrate an alternative model for digital history research projects, one that perhaps can help to provide lasting, sustained benefits for instruction and research in an otherwise dynamic and highly changeable scholarly landscape.

References

[1] Cynthia Lee. "New app helps students learn to read ancient Japanese writing form." UCLA Daily Bruin, Oct. 8, 2015. http://newsroom.ucla.edu/stories/new-app-helps-students-learn-to-read-ancient-japanese-writing-form.

[2] Heizō Kitagawara 北川原平造. "Shikika no kōzō: Kokin wakashū nōto 四季歌の構造: 古今和歌集ノート." Kiyō: Ueda Joshi Tanki Daigaku 紀要：上田女子短期大学. March 31, 1993. http://ci.nii.ac.jp/els/110006406589.pdf?id=ART0008407967&type=pdf&lang=jp&host=cinii&order_no=&ppv_type=0&lang_sw=&no=1462493837&cp=.

[3] Hilofumi Yamamoto. "The Differences of Connotations between Two Flowers, Plum and Cherry, in Classical Japanese Poetry, 10th Century." JADH Conference 2015: Encoding Cultural Resources. September 1-3, 2015, Kyoto University Institute for Research in Humanities: 31-32.

[4] See David Mimno, et al. "The Telltale Hat: LDA and Classification Problems in a Large Folklore Corpus." Digital Humanities 2014. Lausanne, Switzerland, July 10, 2014. http://dharchive.org/paper/DH2014/Paper-163.xml.

go rich :: go minimal

Federico Caria (Cologne University)

This text is inspired by the DHSI workshop on Minimal Computing run by Alex Gil from the GO : DH group, at Columbia University. Minimal Computing can be defined as an approach to computing done under a set of hardware and software constraints, which present more than one analogy with my work in the context of DiXiT. Here I try to develop a sense of what Minimal Computing has to offer to better conceptualize the results of our evaluation. We are conducting focus groups in the context DiXiT, an EU project funded under Marie Curie Action Program, that offer a coordinated training in the multi-disciplinary skills, technologies, theories, and methods of digital scholarly editing. After introducing what Minimal Computing is, I want to draw on how Minimal Computing understands 'usefulness', to discuss the preliminary results of our focus groups.

In illustrating what Minimal Computing is about, Gil (2015) focuses on necessity.

"we prefer to (not) define minimal computing around the question "What do we need?" If we do so, our orientations vis-a-vis ease of use, ease of creation, increased access and reductions in computing—and by extension, electricity— become clearer" (Gil 2015).

Like Minimal Computing, HCI principles and practices are about setting the right constraints. Usefulness is one of these. In particular, understanding and implementing constraints basing on what is useful will potentially aid in usability and help users engage your design with minimal frustration, that is a desired outcome of our testing digital editions.

Digital editions are young media and studying users is often overlooked in the field of scholarly publishing, where editors often design for themselves. On the contrary, we believe that a very good way to reconstruct the relationship between scholars and the socio-technical mechanisms of production and dissemination in the digital realm is studying interaction between users and digital representations moving from observable data. Just like Minimal Computing, studying usefulness is important to resist "the culture of user-friendliness" (Gil ibid.), putting on the foreground the "quality of use" instead of easiness, cause we know that scholars will stick to the laptop to learn the code if they find it useful!

In essence, although our questions arise from a typical design context aimed at correcting skeumorphism, we are also asking "scholars around the world— librarians, professors, students, cultural workers, independent" (Gil ibid.) about their needs and wants. In particular, like Minimal Computing, we gather these data to understand "what is enough" or "what is the finished stairway" (Gil ibid.) of different target users. Our preliminary findings suggest hat humanities scholars, practitioners and students not involved in the digital humanities generally find digital resources extremely useful; in most cases, they just do not know about their existence. According to our interviews, we are pushed to assume that most intended users would warmly welcome the possibility to rely on a digital resource or thematic collection related to their topic. But our data also suggests that there is a threshold under which technology is not perceived as useful anymore by humanities scholars and students, who still largely work with print resources.

Again, usability is about setting thresholds and limits, and usefulness is crucial to adoption and use, which in the networked environment improves the chance of preservation. Sustainability is a concern for us as it is for Minimal Computing. The real crisis of the "crisis of the humanities" is that it's not a crisis but a chronic condition! (Risam 2015). Soon "the natural inclination" of electronic information "to change, to grow, and to finally disappear" will cease to function as an aesthetic conceit and become instead a full-blown cultural crisis (Kirschenbaum 2001). That our scholarly publishing is an unsustainable business seems undeniable: it costs big money to produce an edition, and there is no way to know whether a big and expensive initiative will fall out of use or become unaccessible in a few years' time.

Although the principles of Minimal Computing were turned into a "modelo digital" (see Sustainable Authorship in Plain Text, Tennen and Wythoff, 2014) to answer publishing needs that are typical of Southern media ecologies, I think there is quite a room for a potential application to rich ones. Minimal Computing provides an interesting interpretative framework to understand usefulness as an ethical choice. In the end, if we see, Oroza definition of "arquitectura de la necesidad":

"una arquitectura de la urgencia y la precariedad. Su segunda y más importante función es metafórica: enuncia una arquitectura que es su propio diagrama. La casa deviene una estructura que relaciona, un modelo físico que asocia al individuo necesidades materiales, tecnologías, los límites y posibilidades legales y económicas" (Oroza 1997).
does not fit less those who are about to face precariousness and urgency, let's say, in the medium term.
I wonder if minimalism is something that editors producing in the context of maximalist media ecologies would consider. The question is how to be essential and effective at the same time, in the context of digital scholarly editions out of the postcolonial world? Would a static website ever meet the standards of the many 'monumental' digital editions that can be found in the catalogues? and how can usability help enhancing essentiality and effectiveness? My presentation touches on issues of access, funding, language, standards, and case studies in the attempt to explore the challenges and opportunities in the application of minimalist principles to design for sustainable editions.

References

[1] Drucker, J. (2011) Humanities Approaches to Interface Theory, in Culture Machine, 12. Accessed athttp://www.culturemachine.net/index.php/cm/article/viewarticle/434
[2] Drucker, J. (2014) Graphesis. Visual Forms of Knowledge Production, Harvard University Press
[3] Galloway, A. R. (2012) The Interface Effect, Cambridge.
[4] Fiormonte, D. (2014) Digital Humanities from a Global Perspective, in Laboratorio dell'ISPF, XI, 2014. Accessed at
[5] Gibbs, F., Owens, T. (2012) Building Better Digital Humanities Tools: Toward broader audiences and user-centered designs, (2012) in Digital Humanities Quarterly, 6:2. Accessed athttp://www.digitalhumanities.org/dhq/vol/6/2/000136/000136.html
[6] Gil, A. (2012) The User the Learner and the Machines we Make. Accessed at http://go-dh.github.io/mincomp/thoughts/2015/05/21/user-vs-learner/
[7] Gil, A. (2013) Around DH in 80 Days, accessed at http://www.arounddh.org
[8] Kirschenbaum, M. G. (2001) Materiality and Matter and Stuff: What Electronic Texts Are Made Of, in Electronic Book Review, accessed at http://www.electronicbookreview.com/thread/e!lectropoetics/sited
[9] Koohang, A. (2004) Expanding the Concept of Usability, in Informing Science Journal, 7, 129-41.
[10] Kulesz, O. (2011) Digital Publishing in Developing Countries. International Alliance of Independent Publishers. Accessed at http://alliance-lab.org/etude/wp-content/uploads/digital_publishing.pdf
[11] Oroza, E. (2008) Statement of Necessity, accessed at http://architectureofnecessity.com
[12] Pierazzo, E. (2015) Digital Scholarly Editing. Theories, Models and Methods, Ashgate.
[13] Risam, R. (2015) Across Two (Imperial Cultures). Video of the talk is available here (see ~5:10:00):https://www.youtube.com/watch?v=NeH2QOUf4Qo.
[14] Rucker, S., Radzikowska, M., Sinclair, S. (2011) Visual Interface Design for Digital Cultural Heritage. A Guide to Rich-Prospect Browsing, Farnham, Burlington: Ashgate.
[15] Shirky, C. (1999) An Open Letter to Jakob Nielsen. Accessed at http://www.shirky.com/writings/nielsen.html
[16] Tennen, D, Wythoff, G. (2014) Sustainable Authorship in Plain Text using Pandoc and Markdown, accessed athttp://programminghistorian.org/lessons/sustainable-authorship-in-plain- text-using-pandoc-and-markdown.html
[17] Unsworth, J. (2000) Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this?, part of a symposium on "Humanities Computing: formal methods, experimental practice" sponsored by King's College, London.
[18] Warwick, C., Terras, M., Huntington, P., Pappa, N. (2008) If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities through Statistical Analysis of User Log Data. in LIT LINGUIST COMPUT, 23 (1) 85 - 102
[19] Warwick, C., Terras, M., Nyhan, J. (2012) Digital Humanities in Practice, Facet Publishing, London.

[20] Kirschenbaum, M. G. (2001) Materiality and Matter and Stuff: What Electronic Texts Are Made Of, in Electronic Book Review, accessed at http://www.electronicbookreview.com/thread/electropoetics/sited

Proceedings of JADH Conference, vol. 2016
Published by the Historiographical Institute, The University of Tokyo
3-1, Hongo 3, Bunkyo-ku, Tokyo, Japan
Online edition: ISSN 2432-3144 Print edition: ISSN 2432-3187
Editor: Taizo Yamada