The Japanese Association for Digital Humanities is pleased to announce its second annual conference, to be held at the University of Tokyo, Japan, 15-17 September, 2012. The conference will feature posters, papers and panels. We invite proposals on all aspects of digital humanities internationally, and especially encourage papers treating topics that deal with the ways that humanistic studies are being inherited by humanities scholars in the digital age.

In this decade, the digitization of cultural resources has been carried out extensively by various projects and organizations, taking the ever-growing Internet as the main infrastructure. However, in Japan, such activities have tended to be carried out by practitioners and researchers of information technology—often without sufficient cooperation with humanities scholars. Therefore, one of the aims of JADH2012 is to raise awareness regarding the efforts of humanities researchers and to have some discussion about this area, so that we may provide new support for various approaches toward inheriting humanities in the digital age.

Hosted by: JADH 2012 organizing committee Under the auspices of Japanese Association for Digital Humanities
Co-hosted by: Center for Evolving Humanities, Faculty of Letters, University of Tokyo / Digital Humanities Core, Interfaculty Initiative in Information studies, University of Tokyo / Center for Knowledge Structuring, University of Tokyo / International Institute for Digital Humanities
Co-sponsored by: SIG Computers and the Humanities, Information Processing Society of Japan / Japan Art Documentation Society / Alliance of Digital Humanities Organizations
JADH Conference 2012
“Inheriting Humanities”

15-17 September 2012

Faculty of Engineering Bldg.2 (Hongo Campus), University of Tokyo
於東京大学本郷キャンパス工学部2号館

Hosted by JADH 2012 organizing committee
Under the auspices of Japanese Association for Digital Humanities

Co-hosted by
Center for Evolving Humanities, Graduate School of Humanities and Sociology, University of Tokyo
Digital Humanities Core, Interfaculty Initiative in Information studies, University of Tokyo
Center for Knowledge Structuring, University of Tokyo
International Institute for Digital Humanities

共催
東京大学大学院人文社会系研究科次世代人文学開発センター
東京大学大学院情報学環デジタル・ヒューマニティーズ・コア
東京大学知の構造化センター
一般財団法人人文情報学研究所

Co-sponsored by
Alliance of Digital Humanities Organizations
SIG Computers and the Humanities, Information Processing Society of Japan
Japan Art Documentation Society

協賛
デジタル・ヒューマニティーズ機関連合
情報処理学会人文科学とコンピュータ研究会
アート・ドキュメンテーション学会
Programme Committee:

Hiroyuki Akama (Tokyo Institute of Technology, Japan)
Paul Arthur (Australian National University, Australia)
Neil Fraistat (University of Maryland, USA)
Shoichiro Hara (Kyoto University, Japan), Vice chair
Jieh Hsiang (National Taiwan University, Taiwan)
Mitsuyuki Inaba (Ritsumeikan University, Japan)
Jan Christoph Meister (University of Hamburg, Germany)
Charles Muller (University of Tokyo, Japan), Chair
Hajime Murai (Tokyo Institute of Technology, Japan)
Maki Miyake (Osaka University, Japan)
Kiyonori Nagasaki (International Institute for Digital Humanities, Japan)
Lisa Lena Opas-Hänninen (University of Oulu, Finland)
Espen Ore (University of Oslo, Norway)
Geoffrey Rockwell (University of Alberta, Canada)
Susan Schreibman (Trinity College Dublin, Ireland)
Masahiro Shimoda (University of Tokyo, Japan)
Raymond Siemens (University of Victoria, Canada)
Takafumi Suzuki (Toyo University, Japan)
Tomoji Tabata (Osaka University, Japan), Vice chair
Norio Togiya (University of Tokyo, Japan)
Christian Wittern (Kyoto University, Japan)

Organizing Committee:

Masahiro Shimoda (University of Tokyo), Chair
Hidetaka Ishida (University of Tokyo), Vice chair
Kiyonori Nagasaki (International Institute for Digital Humanities), Secretary
Motoo Endo (University of Tokyo)
Masato Kobayashi (University of Tokyo)
Kenji Sato (University of Tokyo)
Yuki Takinami (University of Tokyo)
Tomoji Tabata (Osaka University)
Toru Tomabechi (International Institute for Digital Humanities)
Takeshi Nakajj (University of Tokyo)
Hiroaki Nagashima (University of Tokyo)
Yusuke Nakamura (University of Tokyo)
Kenryo Minowa (University of Tokyo)
A. Charles Muller (University of Tokyo)
Hideki Mima (University of Tokyo)
Hideyuki Hori (University of Tokyo)
Syunya Yoshimi (University of Tokyo)
Taizo Yamada (National Institutes for the Humanities)
Table of Contents

Plenary Talks

16th Sep. 16:40-17:40, Room A
‘In Dreams Begin Responsibilities’ The Politics and Hermeneutics of DH
Speaker: Susan Schreibman (Trinity College Dublin)
Respondent: Raymond Siemens (University of Victoria)
Chair: Masahiro Shimoda (University of Tokyo)

17th Sep. 9:00-10:10, Room A
Teaching DH: an absurdity or a necessity?
Speaker: Elena Pierazzo (King's College London)
Chair: A. Charles Muller (University of Tokyo)

17th Sep. 13:30-15:00, Room A
Too Small To Fail: the Scholars' Lab at the University of Virginia Library
Speaker: Bethany Nowviskie (University of Virginia)
Chair: Masahiro Shimoda (University of Tokyo)

Sessions
Paper Sessions

A-1: Literary analysis (16th Sep. 10:20-11:50, Room A)
Chair: Lisa-Lena Opas-Hänninen

Network Analysis of Modern Japanese Poetry: Methods and Tools for Exploring Literary Influence and Affiliation
Hoyt Long and Peter Leonard (University of Chicago)
Evaluating Validity of Classic Rhetorical Structure by Comparing with Random Division and Random Combination  
Hajime Murai (Tokyo Institute of Technology)  

Developing a TEI Based Analytical Tool for Textual Variants  
Maki Miyake (Osaka University)  

Text-mining Linguistic Variations from a Diachronic Perspective: An experiment in textometry  
Tomoji Tabata (Osaka University)  

B-1: Scholarly collaboration (16th Sep. 10:20-11:50, Room B)  
Chair: Masahiro Shimoda  

Understanding the Social Edition, Theoretically and Through Iterative Implementation  
Ray Siemens and Constance Crompton (University of Victoria)  

Old Tibetan Documents Online: A Collaborative Editing Project  
Kuninori Matsuda and Izumi Hoshi (Tokyo University of Foreign Studies)  

Description for Restoring Process of “Shoso-in Monjo” using the Shoso-in Monjo Database (SOMODA)  
Makoto Goto (Hanzono University)  

Text Representation and Interchange in the Digital Age  
Christian Wittern (Kyoto University)  

A-2: Analytic method (16th Sep. 13:20-14:40, Room A)  
Chair: Christian Wittern  

Learning of Page Splitting Rule for Reading Order Estimation on Document Analysis  
Makoto Tanji and Hideki Mima (University of Tokyo)  

A Text Analysis Method Using Nonparametric Bayesian Model for Japanese Historical Materials  
Taizo Yamada (National Institutes for the Humanities), Satoshi Inoue, Tamaki Endo and Noriko Kurushima (The University of Tokyo)
SMART-GS system: a software for historians by historians
Susumu Hayashi (Kyoto University), Kenro Aihara (National Institute of Informatics),
Minao Kukita and Makoto Ohura (Kyoto University)

B-2: Time and spatial analysis (16th Sep. 13:20-14:40, Room B)
Chair: Harold Short

Geographical Clustering of Archaeological Site Distributions with R: Using Delaunay
Triangulation
Tomohiko Matsumori (Doshisha University)

The spatial analysis about the Gun-ga distribution in the Ancient Japanese History
Yoichi Seino (Kyoto University)

Sharing Genealogical Spaces for Cultural and Social Anthropological Studies with the
Alliance, a Kinship Database and Genealogy Management System
Shigenobu Sugito (Sugiyama Jogakuen University)

A-3: Human interaction (17th Sep. 10:30-12:00, Room A)
Chair: Tomoji Tabata

With the time-zone differences and the busy schedules, setting up a cross-atlantic
conference call takes some time: Benefits, challenges, and strategies to support interna-
tional research teams
Lynne Siemens (University of Victoria)

The Importance of Storytelling in a Digital World: Or, What the Digital Humanities can
learn from TED
Ryan Hunt (University of York)

Constructing 3D Metaverse for Inheriting and Sharing Japanese Traditional Culture
Mitsuyuki Inaba (Ritsumeikan University)

B-3: Database usage (17th Sep. 10:30-12:00, Room B)
Chair: Espen S. Ore

DH and Disaster Management: An Overview of the UC CEISMIC Digital Archive
James Smithies (University of Canterbury)
Publishing manuscript catalogues online, an attempt by the Nepalese-German Manuscript Cataloguing Project
Kengo Harimoto (University of Hamburg) 33

Australian Indigenous Languages Database, AUSTLANG
Kazuko Obata (Australian Institute of Aboriginal and Torres Strait Islander Studies) 35

A-4: Database construction (17th Sep. 15:20-16:50, Room A)
Chair: Takafumi Suzuki

The Orationes Project: Bringing a Restoration Manuscript Online
Ilkka Juuso, Tuomo Toljamo, Lisa Lena Opas-Hänninen, Tapio Seppänen and Anthony Johnson (Oulu University) 36

Design of Organizing MANGA Multimedia
Shoichiro Hara (Kyoto University) and Motomu Naito (Knowledge Synergy Inc.) 37

Corpus Sharing Strategy for Descriptive Linguistics
Kazushi Ohya (Tsurumi University) 39

B-4: Education (17th Sep. 15:20-16:50, Room B)

Panel: Undergraduate Models for Inheriting Interdisciplinary and Humanistic Studies in the Digital Age
Alex Benkhart, Janet Thomas Simons, Kyoko Omori and Angel David Nieves (Hamilton College) 41

Poster/Demo Session

C-1: Poster/Demo Session (16th Sep. 15:00-16:20, Room C)
Chair: Charles Muller

Posters

Quantitative analysis of style change and conversational sentence within the works of contemporary novel writer
Akira Kudo (Tokyo Institute of Technology) 43

Logical Layout Analysis for Japanese Journal in Humanities with Machine Learning
Katsuya Masuda and Hideki Mima (University of Tokyo) 45
Stylistic analysis of tweets that are likely to be shared
Yui Arakawa (University of Tsukuba), Akihiro Kameda (University of Tokyo), Akiko Aizawa (National Institute of Informatics) and Takafumi Suzuki (Toyo University)

On a Digital Database of Ancient Egyptian Hieratic Palaeography
Masakatsu Nagai (University of Tsukuba)

Design of Serial Comparison Model for the Diachronic Corpus Study of Japanese
Hilofumi Yamamoto (Tokyo Institute of Technology / University of California, San Diego), Makiro Tanaka (National Institute of Japanese Language and Linguistics) and Yasu-Hiro Kondo (Aoyama Gakuin University / National Institute of Japanese Language and Linguistics)

Emotive Adjectives and Verbs of the Heian Japanese
Makiro Tanaka (National Institute for Japanese Language and Linguistics) and Hilofumi Yamamoto (Tokyo Institute of Technology)

Toward Non-Consumptive Formal Evaluation Challenges Using the HathiTrust Research Center Digital Collections
J. Stephen Downie (University of Illinois Urbana-Champaign), Beth Plale (Indiana University), Robert H. McDonald (Indiana University) and Marshall Scott Poole (University of Illinois Urbana-Champaign)

Interactive display of shining gilded illumination in Keio Gutenberg bible
Hiroyo Ishikawa, Fumihiko Kamemura, Kunitake Kaneko, Hideo Saito and Takami Matsuda (Keio University)

[AInvited poster]
A Quantitative Investigation of English Accents
Martijn Wieling (University of Groningen and University of Tübingen), Jelke Bloem (University of Groningen), Kaitlin Mignella (University of Groningen), Mona Timmermeister (University of Groningen), R. Harald Baayen (University of Tübingen), and John Nerbonne (University of Groningen and Freiburg Institute of Advanced Studies)

Demonstrations

Australian Indigenous Languages Database, AUSTLANG
Kazuko Obata (Australian Institute of Aboriginal and Torres Strait Islander Studies)

Constructing 3D Metaverse for Inheriting and Sharing Japanese Traditional Culture
Mitsuyuki Inaba (Ritsumeikan University)
Publishing manuscript catalogues online, an attempt by the Nepalese-German Manuscript Cataloguing Project
*Kengo Harimoto (University of Hamburg)*

SMART-GS system: a software for historians by historians
*Susumu Hayashi (Kyoto University), Kenro Aihara (National Institute of Informatics),
Minao Kukita and Makoto Ohura (Kyoto University)*

The spatial analysis about the Gun-ga distribution in the Ancient Japanese History.
*Yoichi Seino (Kyoto University)*

---

**Workshop**
*(15th Sep. 9:00-17:00)*

Introduction to Document-based and Genetic editing in TEI
Lecturer: *Elena Pierazzo (King's College London)*

---

**Pre-conference Lecture**
*(15th Sep. 17:30-19:00)*

HathiTrust Research Center: Pushing the Frontiers of Large Scale Text Analytics
Speaker: *J. Stephen Downie (University of Illinois Urbana-Champaign)*
Chair: *Shunya Yoshimi (University of Tokyo)*

---

x
[Plenary Talk]
‘In Dreams Begin Responsibilities’ The Politics and Hermeneutics of DH

Susan Schreibman (Trinity College Dublin)

Digital Humanities is being increasingly noticed outside the relatively small group of researchers who have self-identified with the field over the past several decades. It is being noticed in the popular press, by our colleagues, granting agencies, and at the highest levels of university administration.

With this growing recognition it is also becoming a more politicised space, with questions increasingly raised about who and what is at the centre of DH, and what at the periphery; how funding is disbursed, and what the role of DH within a larger humanities context is — a saviour of the humanities or the last nail in the coffin.

This talk will explore these issues, focusing on contemporary debates as well as the results of a community-based exercise that the editors of the Companion to Digital Humanities have undertaken this summer to preparation for the second edition.
[Plenary Talk]
Teaching DH: an absurdity or a necessity?

_Elena Pierazzo (King's College London)_

As many universities around the globe are designing or thinking about designing academic courses on some flavour of Digital Humanities, some questions about the lifespan of such courses and their content have been raised, particularly given that many courses in this discipline that started in the early 2000s have closed already. At King’s College London we are now about to reach the milestone of ten years of some of our DH courses, but the balance of the past decade is not all positive and shades about our uncertain future are spreading. King’s College London under the guidance of Harold Short has been one of the first Universities worldwide to offer structured programmes in DH at undergraduate and postgraduate level, and was also the first to offer a doctoral programme. The advantage of being pioneering in new disciplines and fields are that one can set new standards and model the field in new and interesting ways; the disadvantages are that one cannot learn from someone else’s mistakes. There were mistakes – we made mistakes which lead to the closure of the Undergraduate programme in 2008 due to the very low number of students – but there were also successes: the MA and the PhD programmes are thriving and we are now facing more student applications that we are able to accept. Furthermore, a new climate in the Digital Humanities and a growing interest in the impact of digital technology in everyday life are leading to the design of a new Undergraduate programme, the content of which will be considerably different from the first (unsuccessful) programme. In fact the UG programme will not be about DH at all, but will concentrate on Digital Culture and Society.

The tale of the mixed success of the teaching programmes at King’s College London will be used to discuss fundamental issues of our discipline such as: are Digital Humanities a topic that can be taught at Undergraduate level at all? To learn DH, do students need some prior competence in a Humanities topic? If so, then is DH exclusively research-oriented or can we envisage some academic curriculum which can lead to successful employment outside Academia? And what constitutes a programme in Digital Humanities? What should we include in our syllabi?

If it is true that a discipline is not a discipline without a teaching programme and that a teaching programme defines the content of a discipline, then answering these questions is of crucial importance to define what Digital Humanities are and what constitutes their intellectual agenda.
[Plenary Talk]
Too Small To Fail: the Scholars' Lab at the University of Virginia Library

Bethany Nowviskie (University of Virginia)

How can a service-oriented digital center maintain an independent research profile? How can a tiny team make a big impact? How can an interdisciplinary community stay brave, and light-hearted, and focused — and set the best possible example for the next generation of digital humanities scholars and professionals? In 2006, the University of Virginia combined its internationally-known Electronic Text Center (established in 1992) with two other longstanding organizations: the Library's Geospatial and Statistical Data Center and a UVa Center for Research Computing Support. The following year, a new team was added: software developers who would become a division for Digital Scholarship Research and Development. The whole (still only a group of a dozen full-time collaborators and affiliated graduate students) became known as the "Scholars' Lab." Among its other roles at the University of Virginia Library, the Scholars' Lab supports and inspires faculty and student scholarship in the digital humanities, fosters intellectual community at home and abroad, and undertakes unique research and teaching projects. Most of all (in an age of wild expansion), it tries to stay "too small to fail." The SLab's director, Bethany Nowviskie, will discuss what this philosophy means to her faculty, students, and staff: how smallness in all of its varieties can become a virtue and an enabler of audacious thinking — and how the Scholars' Lab works to create a space in which little risks are taken and little failures celebrated every day.

Hoyt Long and Peter Leonard (University of Chicago)

In the past few years, social network analysis and corresponding visualization tools have opened up vast new possibilities for how we approach literary history, particularly questions about influence and affiliation. “Graphs” of the linkages between authors, texts, publications and similar data have proved useful not only for computational analysis of large-scale social data, but also for interactive and exploratory interfaces that lead to a more productive and rigorous dialogue with the literary archive. Our presentation demonstrates some of the advantages of these new tools and methods as applied to the case of modern Japanese poetry in the first half of the twentieth century.

There is a well-established tradition in Japanese literary scholarship of producing richly detailed, but ultimately static, bibliographic data. Network analysis and visualization transforms this material into a dynamic site of exploration and discovery. We demonstrate this with a prototype based on an index of over 100,000 poems and essays published in 166 avant-garde poetry journals by nearly 4,000 poets between the years 1920 and 1944. This was a period of rapid flux in the evolution of modern Japanese poetry, characterized by the continuous grouping and regrouping of poets around small-scale journals organized through mutual acquaintance and/or common aesthetic and ideological purpose. These (re)groupings constitute a dataset perfect for network analysis and visualization. Our prototype aims to discover patterns of affiliation at “macroscopic” scale, and ask how these patterns reinforce or complicate previous studies of individuals and literary movements in isolation.

We first describe how to transform linear publication histories into dynamic, edge-and-node network data and explain how the mathematical properties of the resulting network graphs (e.g., centrality measures, clustering coefficients, homophily) can be used to generate hypotheses about the evolving structure of social relations between poets and various poetics styles over time. We then discuss how this same network data might be shared worldwide, constituting a globally-distributed research corpus. We suggest that the combination of graph databases and browser-based visualization engines allows for investigation and manipulation of this rich data source by scholars, students, and the general public. While specialized analysis and general data exploration currently exist as distinct use cases, we believe that the two will eventually converge as the tools improve and become further integrated with web-based interfaces.

In this sense, our project is an example of how the digitization of humanities-style data can be carried out both for, and by, humanists. And in ways that promote collaboration and the shared curation of data across disciplinary and geographical divides.

<Keywords: Poetry, Literary History, Network Analysis, Visualization, Data Mining>
[A-1] Evaluating Validity of Classic Rhetorical Structure by Comparing with Random Division and Random Combination

Hajime Murai (Tokyo Institute of Technology)

In order to interpret the Bible precisely, literary criticism is a promising field. It is a methodology for analyzing the Bible as literature and its use of literary techniques. A marked literary characteristic of the Bible is its sophisticated structures, which comprise classic rhetorical structures such as chiasmus, concentric structures, and parallelisms.

There are several merits to identifying rhetorical structures in the Bible. It can clarify the divisions in a text; moreover, correspondences of phrases in them signify deeper interpretation. If the rhetorical structure is the concentric structure, the main theme of that text is also clarified.

However, there are some problems regarding rhetorical structures. First, there is no clear definition regarding what is a valid correspondence. Some structures correspond by words or phrases, but more abstract themes may also be the element of correspondences. The length of the text unit is not uniform. Some structures are composed of phrases while other structures are composed of pericopes. Therefore, a quantitative validation method for the rhetorical structure of the Bible is necessary.

In this methodology, first, the relationships between each part of the text in the rhetorical structure were validated on the basis of the common occurrence of rare words and phrases. If corresponding pairs of pericopes more frequently include common rare words and phrases, the probability of intentional arrangement is believed to be higher.

Second, on the basis of the mean and standard deviation of a random combination of pericopes, the probability of accidental occurrences of common words and phrases in the test hypothesis is calculated. In this research, two types of random patterns are evaluated. One is random combination of pericopes, and the other is random division of pericopes. The distribution of numbers of common words and phrases is assumed to be a normal distribution. The results are utilized in order to determine whether it exceeds the level of statistical significance. The occurrence of words and phrases were counted in the forms of one word, a two-word phrase, a three-word phrase, and a five-word window; 10 percent, 20 percent, and 30 percent were set as the threshold for each type of form. “Parallel Concentric Structures within the Bible” was selected as a hypothesis for validation.

As a result, in many texts of the Bible it is found that the hypothesis is validated in the relationships of words that occurred in less than 20 percent of part of the text.

A similar tendency in statistical significance among various books of the Bible seems to confirm that a unified system of rhetorical structure is included in those texts.

<Keywords: Rhetorical Structure, Bible, Chiasm, Quantitative Analysis>
[A-1] Developing a TEI Based Analytical Tool for Textual Variants

Maki Miyake (Osaka University)

In this study, we present ongoing research on computational analysis of text variants. More specifically, we focus on investigating the similarities and differences among the influential modern editions of the Greek New Testament. In order to elucidate between the editions, the study proposes a user-friendly analytical tool based on the Text Encoding Initiative (TEI) Guidelines.

The latest edition (the 27th) of the Nestle-Aland Greek New Testament was used for the study, for this series of editions is considered to be one of the most useful critical materials for biblical studies. Although a rigorous survey of the textual differences from the most important modern editions has been conducted by the editors, the result are shown in the form of reference to the critical apparatus indicated by critical signs in the appendix section. It is quite difficult for those who are not experts in textual criticism to identify these differences all at once.

As for the pilot study, we focused on two modern editions such as Westcott-Hort (1881) and Scrivener (1894) within the synoptic Gospels; the Gospels of Matthew, Mark and Luke are similar in structure, content, and wording. We applied a text segmentation algorithm called TextTiling to computing the similarities in comparison with the standard scholarly edition of the Nestle-Aland. TextTiling proposed by Marti Hearst (1994) is a method for automatically subdividing texts based on the cosine similarity measure into multi-paragraph units that represent passages, or subtopics. The statistical results clearly represent segments of the similarities and differences between a given edition and Nestle-Aland.

Simultaneously, we are developing the analytical tool, capable to compare the textual variants combining the statistical results and the digital critical edition. For the purpose of providing digital critical editions with high usability, we attempted to implement the TEI-encoding scheme using the"critical apparatus module" in digital critical editions based on the Nestle-Aland edition.

<Keywords: Text Mining, Vector Space Model, Text Encoding Initiative>
[A-1] Text-mining Linguistic Variations from a Diachronic Perspective: An experiment in textometry

Tomoji Tabata (Osaka University)

The aim of the present paper is to investigate linguistic variations in the Century of Prose Corpus (COPC) Part B. The COPC, compiled by Louis T. Milic, covers the time period of 1680–1780 and provides an empirical basis for the study of prose style in the era. The corpus is made up of two parts. Part A comprises texts written by 20 major prose writers of the period such as Dryden, Addison, Swift, and Johnson. Each writer is represented by three 5,000-word text samples. Part B, on the other hand, is composed of texts written by 100 second-tier writers. What is of special interest about this Part B is that the 100 texts can further be divided into ten subsets of ten texts, ten in each of ten registers and ten in each decade of the century. The COPC Part B can therefore be seen as a multi-register corpus with a diachronic perspective.

Although the COPC has been referenced in a number of diachronic researches (Rissanen, Kytö & Heikkonen 1997a/b, McKenny 2003, Yáñez-Bouza 2006, and Mizuno 2006), few studies have investigated Part B of the corpus per se. Notable exceptions are Mizuno (1995) and Milic and Slane (1994). Mizuno tried to find extremely rare syntactic phenomena such as ‘demonstrative + possessive pronoun constructions’ and ‘push-down relatives’. Milic and Slane attempted to make a quantitative description of genres (registers) in Part B and claimed that it would be necessary to employ a more sophisticated statistical technique than what they used, in order to extract features that would distinguish a genre from a collection of other cases. Despite their claim, however, subsequent investigations have yet to appear.

The present study employs a multivariate approach demonstrated in Tabata (2004) to observe the structure of the COPC Part B from a macroscopic standpoint. The particular focus is placed on 100 most common words across 100 texts. The 100 most common words account for 52.7% of total word-tokens in the corpus. A multivariate analysis of these linguistic variables makes it possible to visualize complex interrelationships among linguistic variables, interrelationships among texts, and interrelationships between linguistic variables and texts. By the use of this methodology, the author will try to illustrate dimensions of linguistic variations across registers as well as linguistic variations over time in the century of prose.

References


— (1997b) Grammaticalization at Work: Studies of Long-Term Developments in English, Mouton de Gruyter.


<Keywords: variation, register, Century of Prose corpus, stylometry, diachronic textometry>

Ray Siemens and Constance Crompton (University of Victoria)

In the very early days of the world wide web, but well into a period in which our community understood the positive and transformative impact that computational technique has had on scholarly editing, Fortier (1991) reminded us that literary studies is and always has been focused on the engagement of texts regardless of interpretive theoretical predisposition. In digital literary studies, that textual focus manifests in a number of theories about the nature of the text in general and the electronic scholarly edition in particular, and has developed such that we can begin to construct, in relatively straightforward manner, a basic typology of electronic scholarly editions via the approach each type takes in handling and engaging its textual materials: from edited electronic text plus analytical tools for its readers (dynamic text), to text plus a static set of additional supporting materials in digital form for reader navigation and subsequent analysis (hypertextual edition), to text augmented by both dynamic analytical means and hypertextually-linked access to fixed resources plus automated means of discovering and interrelating external resources (dynamic edition). Such a typology, reductive as it may be, allows us to look forward - as Robinson (2010), Shillingsburg (2006), Bryant (2002), McGann (2001), Gabler (2010) and many others have encouraged us to do, variously - to what lies ahead in our treatment of the text, and the textual editions, that lie at the core of our contemplation in literary studies and similar disciplines.

Well into what is often called the age of Web 2.0 - becoming immersed as we are in a generation of online tools facilitating collaboration, information sharing, and interoperability; immersed as we are by social media interaction on the web - it is worth noting that the types of electronic scholarly editions we see prominently today were largely developed before the ubiquity of the web that we now enjoy and do not accurately reflect the full range of useful possibilities present for academic engagement and interaction around the textual materials that are our focus. While the electronic medium is most certainly a productive space in which to present and analyse editions, it is increasingly more difficult to ignore the influence of new and emerging possibilities for the electronic scholarly edition in the current phase in the social formation of the web. As such, our understanding of the electronic scholarly edition in its current form requires reconsideration in light of the collaborative potential of already extant and newly-emerging digital technologies; put another way, we need to extend our understanding of the scholarly edition in light of new models of edition production that embrace social networking and its commensurate tools.

This paper offers an understanding the scholarly edition in the context of new and emerging social media from two pertinent perspectives: the first from the perspective of its theoretical context, particularly as that context intersects with a utility-based consideration
of the toolkit that allows us to consider the social edition as an extension of the traditions in which it is situated and which it has the potential to inform productively; the second is from the perspective of an iterative implementation of one such edition, carried out via a research team operating in conjunction with an advisory group representing key expertise in the methods and content-area embraced by the edition. The implementation that will form the basis of the example is an edition of the Devonshire MS (BL Add Ms 17492), the first sustained example of men and women writing together in the English literary tradition; see http://en.wikibooks.org/wiki/The_Devonshire_Manuscript.

<Keywords: scholarly editing, editorial theory, social networking, social media, community>
Old Tibetan Documents Online (OTDO) is a corpus of selected Old Tibetan texts (7th to 12th centuries) comprising Dunhuang manuscripts, inscriptions, and related materials. In order to make these documents widely and freely available, the editors cooperated with the Research Institute for Languages and Cultures of Asia and Africa (ILCAA) to transliterate certain texts and publish them on the OTDO website (http://otdo.aa.tufs.ac.jp/) using the KWIC search engine. Additionally, a monograph series was published by the ILCAA.

The PukiWiki-based system was used to collaboratively prepare the texts and related information. The wiki-based system consists of mainly two parts: texts and bibliography related to each corpus. The former comprises text information and text edition sections. The text information section contains catalog numbers, brief explanations, references to the bibliography, links to related images published by the International Dunhuang Project, etc. The text edition section contains the aforementioned transliterated text with annotations. Since the project's initiation, the wiki-based system was found to be quite effective for this type of collaborative work.

However, as the corpus expanded, certain problems or unsatisfactory points in the wiki-based system were observed. For example, a lot of human attention is required to ensure consistency between catalogue and bibliography data, and this leads to unnecessary confusion.

To alleviate these problems, we are developing another collaborative editing system, which will use an AJAX-based interface and PostgreSQL as a backend database. Analogous to the wiki-based system, this system will comprise mainly three parts (i.e., text information, text edition, and bibliography), provide measures for cross-referencing, etc. Naturally, this system will take into account the wiki-based system's features (e.g., the edit history feature).

In this presentation, we report the outline of our project, problems we encountered with the wiki-based system, and merits of the new system and areas for improvement.

<Keywords: Old Tibetan Documents Online (OTDO), collaborative editing, tibetan corpus, database>
[B-1] Description for Restoring Process of “Shoso-in Monjo” using the Shoso-in Monjo Database (SOMODA)

Makoto Goto (Hanazono University)

The Shoso-in Monjo (document) is the generic name for a group of documents that has been passed down through the ages within the Shoso-in of Todaiji Temple. Ever since Tadatomo Hoida extracted official parts of the document and edited the “Seishu” collection by focusing on the first side of the sheet during the Tempo years (1833-1836), “Zokushu,” “Zokushi-Koushu,” “Zokushu-Besshu,” “Zinkai,” and “Zokuzokushu” were continuously compiled during the Meiji era. As a result, the “original” form of the document group has been transformed (45 volumes of “Seishu,” 50 volumes of “Zokushu-Besshu,” 43 volumes of “Zokushi-Koushu,” 50 volumes of “Zokushu-Besshu,” 440 volumes and 2 books of “Zokuzokushu,” and 39 volumes and 3 books of “Zinkai”). Therefore, these “rearranged” documents need to be returned to their original form.

Restoration of the Shoso-in Monjo is necessary for the following reasons. As mentioned above, both rearrangements—those in the Tempo years undertaken by Tadatomo Hoida, as well as those in the Meiji era—were focused on the first side of the sheet. However, in the “Seishu” or “Zokushu” scrolls of these rearrangements, the appropriate connecting and/or binding links were lost because only excellent documents were collected by making arbitrary selections. In addition, the relation between some of the inside and outside parts of sheets is not clear because the rearrangements focused only on the first side of sheet. This relation should be restored.

Therefore, we created an XML database for the Shoso-in Monjo and attempted to restore the document to its original state using a computer. We refer to the database as the ShOsoin MONjo DAtabase (SOMODA).

“SOMODA” is characterized by three features. First, physical and logical structures can be extracted as well as restructured using XSL/XSLT from the XML document, which is an intermediate expression generated using structuring rules. In other words, these rules are considered to generalize or model the process of restoration. By identifying exceptions to these structuring rules, events that had only been recognized vaguely thus far become clear. Until now, researchers have been working without any clear awareness of the structuring rules. Therefore, these efforts may contribute to the development of a method for the “theoretical study on Ancient Japanese Documents.” We don’t apply TEI for “SOMODA”. Because of Shoso-in Monjo is a very complicated text. TEI is not turn to a Shoso-in Monjo.

Second, “restored tables” can easily be created from the XML document, thus greatly reducing the researcher’s efforts. For creating restored tables, descriptions regarding connecting/binding links (presence or absence of tearing or ink marks) and conditions of the sheet are indispensable. These descriptions are yet to be provided;
however, this is a technical issue that is likely to be resolved immediately.

Third, “SOMODA” has prepared an environment in which researchers can contribute by using logical expressions. Digitization of the Shoso-in document, with its numerous difficulties in structure and form, has been considered a complicated issue. However, we have shown that it can be done by using XML/XSLT in historical studies. Improvements toward a more convenient system will be a subject for future studies.

<Keywords: Ancient Japanese Documents, XML Database, Digital Document Restoration>
[B-1] Text Representation and Interchange in the Digital Age

*Christian Wittern (Kyoto University)*

In digital form, text has no 'natural' representation that is dictated by physical constrains of the medium. On the contrary, the digital medium requires the development of an explicit architecture for text. One fundamental distinction that has been made here is that between encoded text and text as a graphical representation. Another useful distinction I would like to consider here is the distinction between file-oriented and database-oriented text representation.

File-oriented text representation in one of its many forms is currently the major form of textual representation, be it as simple as 'plain' text with only minimal formatting or elaborate XML markup or any of the many existing proprietary text formats for word processing.

Database-oriented text representation is less visible and currently in most cases derived from some form of file-oriented representation, such as an index generated for fast full-text searches, but even the print age knew its form of such 'database'-oriented text in the form of concordances and back-of-the-book indices.

All types of text and indeed everything that is represented digitally requires interpretation of the bits used to represent the information in order to present it to the reader or for other purposes. It is thus necessary to consider the represented text and the system used to render it *together* as the equivalent of pre-digital text printed on physical media.

Some file-oriented textual formats, such as plain text or HTML can be rendered meaningfully by a large number of applications, while others, such as Word documents are more specialized and not only require a specific application, but in many cases also rely on a specific version of such a software, a specific version of a specific operating system and might even need additional fonts to be rendered correctly.

Database-oriented texts however, are rarely meaningfully accessed outside of the context they have been generated in and are thus rarely considered a specific form of a text, but rather part of a search application or content management system. Examples include the index of the World Wide Web maintained by companies such as Google, websites such as Wikipedia, or even, for the most digital editions of texts that are represented through an elaborate and idiosyncratic web interface.

In most cases, database-oriented text are generated algorithmically from file-based representations of a text and during the processing information irrelevant to the intended task is thrown away like punctuation, paragraph boundaries or page numbers, other information might be normalized like spelling or case. For this reason, such views on a text are seldom discussed in the context of digital text representation.
In this paper, I will argue that text in a digital environment is more meaningfully understood and processed by seeing different representations of the text as different views on a text that might have different degrees of fidelity to a specific text witness as it originally appeared in print, but serve their own purpose to enhance the total view on the text presented by the sum of these different views. This argument is intended to open up a path to a new environment, which will allow such different views of a text, be they multiple file-based representations in different formats, such as XML, plain text, but also PDF and ePub, the AZW (or mobi) format used by amazon.com or database-oriented formats presented in a browser or a list of search results to coexist in a meaningful way as part of a whole ecosystem of texts, where, for example, references, annotations, bookmarks and so on can easily be carried from one representation of a text to another based on the current need of a user and even be shared among users or published.

<Keywords: Text encoding, text representation, scholarly collaboration>
[A-2] Learning of Page Splitting Rule for Reading Order Estimation on Document Analysis

Makoto Tanji and Hideki Mima (University of Tokyo)

This study presents an approach of Structuring Documents for historical documents (old books, articles, etc.), especially proposes a method of reading order estimation which is more desirable for further development of Digital Humanities.

To analyze undigitalized documents, we need first scan all the document pages, then process OCR (Optical Character Recognition) recognition. Simple OCR process can only give batch of text blocks which are separated into one-by-one page. Although, the types and ordering of text block in a document should be recognized for data accessibility (e.g. speech synthesis and further document analysis).

The authors are proceeding a project of automatic document structuring. It processes a large volume of Japanese historical documents then generates structured data which is integrated to unit of each paper, tagged attributes of author and title information. This paper reports a result of Reading Order Estimation of document blocks as part of the project.

Although there are some researches of reading order estimation, it is difficult to create a rule that adapts various types of all documents. Our method learns ordering rule using an optimization method borrowed from machine learning techniques. It extracts domain specific knowledge (number of page columns, format styles, etc.) and uses these features to split and ordering pages. The obtained ordering rule is represented by a weight vector of these features.

In our experiment study on Japanese historical document "Shiso(Thought)" published by Iwanami Shoten, the parameters of ordering rule were optimized using evolutionary computation. Then it was tested on test set. Our method generated sequence of text blocks with partial order relations with about 0.04 of error (Spearman Distance). It allows us to recognize connection of text blocks and to access structured units of paper. The method can be used for other types of documents. We will present experimental result on other database (e.g. MediaTeam Oulu Document Database).

The purpose of the project is to generate the automatic method to analyze historical documents and understand each document and relationship of inter-documents. This work shows difficulty of dealing with variety types of document by one rule and illuminates effectivity of adaptation for each different types of document, especially frequency map of block boundaries and number of columns.

<Keywords: Document Analysis, Reading Order Estimation, Digital Archiving>

Taizo Yamada (National Institutes for the Humanities), Satoshi Inoue, Tamaki Endo and Noriko Kurushima (University of Tokyo)

1. Introduction

A Research of Japanese history has been a revolutionary advance by some benefits of DH (Digital Humanities). Since texts of Japanese historical materials (such as historical documents, old diaries, and so on) have been encoded (e.g. plain, XML-based, TEI-based) and published via several web systems (e.g. SHIPS\(^1\), nihuINT\(^2\)), researchers of Japanese history can search and obtain research resources such as a text, a catalogue, and an image. However there are hardly systems which provide tagged texts.

In this paper we present an analysis method of texts of Japanese historical materials. Since there is no method of a morphological analysis for a Japanese historical text, we extract some frequent subsequences from untagged texts to use n-gram language model. We consider that the subsequences are meaningful in understanding the texts and the texts can be characterized by the subsequences. Feature vectors of the texts in a vector space model are consisted of the subsequences. To measure similarities between the materials by the vector space, we attempt to gather a material which is similar to the material.

2. Extracting subsequences

We attempt to extract subsequences from a text of a historical material using n-gram language model.

A typical n-gram language model approximates the distribution over sentences using the conditional distribution of each word or character given a context consisting of only the previous \((n - 1)\) words or characters,

\[
P \left( w_i \cup w_{i-1}^{i-1} \right) = P \left( w_i \cup w_{i-n+1}^{i-1} \right), \quad (1)
\]

where \( w \) is a sentence. The probability is typically computed by the maximum likelihood estimate. The language model is simply and powerful because the model can be used without any special dictionaries or thesauruses. However, n-gram model has a zero frequency problem or a sparseness problem, a smoothing technique should be introduced.

Modified Kneser-Ney Smoothing (MKNS) which is a mixture technique of absolute discounting and back off is one of best smoothing technique among heuristic smoothing techniques [1]. However, the smoothing technique cannot generate n-gram probabilistic distributions hierarchically for each degree (n-1). The language model called

\(^1\)http://wwwap.hi.u-tokyo.ac.jp/ships/shipscontroller
\(^2\)http://int.nihu.jp/
Hierarchical Pitman-Yor Language Model (HPYLM) which is introduced in [2] can generate (n-1) degree probabilistic distributions by nonparametric Bayesian model. Using the advantage of HPYLM, variable order hierarchical Pitman-Yor language model (VPYLM) which is a kind of a variable length n-gram model is introduced in [3]. We may expect that using VPYLM some subsequences which characterize the text can be obtained without any dictionaries. In VPYLM, the predictive probability of a symbol $s_t$ in context $h=s_{t-n}, \ldots, s_{t-1}$ is recursively computed by

$$p(s_t|h) = \frac{c(s_t|h) - d \cdot t_{hs} + \theta + d \cdot t_h}{\theta + c |h|} p(s_t|h'), \quad (2)$$

where $\theta$ and $d$ are hyper-parameters in HPYLM, $t_{hs}$ is the seating arrangement parameter in HPYLM, $t_h = \sum_t t_{hs}$, $h' = s_{t-n+1}, \ldots, s_{t-1}$, $c(s_t|h)$ is the count of $s_t$ at $h$. A length of Each n-gram is according to the probability of follows:

$$n_t \sim p(n_t | s, z_{-t}, n_{-t}), \quad (3)$$

$$\propto p(s_t \cup s_{-t}, z_{-t}, n) p(n_t \cup s_{-t}, z_{-t}, n_{-t}), \quad (4)$$

where $n = n_1, n_2, \ldots, n_T$ corresponds to each Markov order ($n=0 \ldots \infty$) from which each symbol $s_t$ originated, $z$ is a vector of a seating parameter in NPYLM, $s_{-t}$ is $s$ excluding $s_t$, $z_{-t}$ is $z$ excluding $z_t$ and $n_{-t}$ is $n$ excluding $z_t$. Because the first term in (4) is the probability of $s_t$ under HPYLM, the probability is given by (2). The second term in (4) is the conditional probability of $n = l$, according to follows:

$$p(n=l|h) = q_l \prod_{i=0}^{l-1} (1-q_i), \quad (5)$$

where $q_i$ is the beta distribution.

To estimate the parameter $n_t$, a Gibbs sampling according to (3) should be executed. The Gibbs sampling is repeat the following procedure: reduction each n-gram count, resampling $n_t$ according to (3), and addition n-gram according to new $n_t$.

Figure 1 shows an example of subsequence extraction using VPYLM from “吉川家文書 右田弘直打渡状”. Certainty there are needless subsequences, but some useful subsequences such as a personal name (“吉川”, “駿河守”), a place name (“石見国”), a date(“至徳二年”) are extracted without any dictionaries.
3. To measure similarities between materials

In the vector space model, we construct a feature vector according to a term which is a subsequence extracted by VYLM from the texts. The vector of material $m_x$ consists of the term frequency $tf(t_i, m_x)$ where $t_i$ is a term. Using the inverse document frequency of $t_i$, the weight of $t_i$ in $m_x$ is calculated as follows:

$$v(i, x) = \text{weight}(t_i, m_x) = \frac{tf(t_i, m_x) \log N}{df_i},$$

where $N$ is a total of the materials, $df_i$ is the document frequency of $t_i$. The element $v(i, x)$ in the feature vector of $m_x$ is the weight $(t_i, m_x)$.

The similarity between $m_x$ and $m_y$ is calculated using the cosine similarity:

$$\text{sim}(m_x, m_y) = \frac{\sum_i v(i, x) v(i, y)}{\sqrt{\sum_i v(i, x)^2 \sum_i v(i, y)^2}},$$

Figure 2 shows an example of similarity between materials where similar materials similar to “鹿王院文書 足利義満御内書” are ranked be (7) and are displayed.
Figure 2: An example of similar documents

4. Conclusion

We attempt to extract several meaningful subsequences automatically from texts of Japanese historical materials, we use variable order hierarchical Pitman-Yor language model (VPYLM). In the point of the precision, the performance of VPYLM may not good, but using a method of a semi-supervised learning (e.g. with a dictionary of a person name), the performance may be better.

Acknowledgment

This work was supported by Grant-in-Aid for Scientific Research (A) (23240031).

Reference


<Keywords: text analysis, nonparametric bayesian model, n-gram, Japanese historical material>

Susumu Hayashi (Kyoto University), Kenro Aihara (National Institute of Informatics), Minao Kukita and Makoto Ohura (Kyoto University)

The SMART-GS system is a tool for transcription of and studies on handwritten documents. It has been successfully applied to three studies in modern history. In these studies, it made transcription work and document studies much easier and less time-consuming. In a case, it enabled to transcribe and study documents handwritten by Hajime Tanabe, a renowned Japanese philosopher of Kyoto School, which had not been able to "decipher" for nearly 50 years. The past, present and futures of the SMART-GS system will be presented.

SMART-GS has the following features:
(1) markups for imaged documents as well as texts,
(2) one to many bilateral-links between markups,
(3) an imaged document may have line information,
(4) words and phrases of imaged documents are searchable,
(5) features for "collaborative transcription."

There are several similar tools and web-services. The genesis of SMART-GS would be unique among them so that the system has been mainly designed and developed by a historian for his and his colleague's historical studies. Due to the development style, there is no communication barriers between historian-users and the developer. This made the system practical rather than experimental, e.g. no experimental technologies except an imaged-document search engine are used. The development style would be the main reason of successful applications.

SMART-GS system was originally developed by the first author Hayashi for his study of history of thoughts on German 19th century mathematics. He was a software engineering professor before he was turned to a historian of thoughts. SMART-GS was developed to support his own study of 19th century German handwritten texts by a prominent mathematician David Hilbert. The system was quite useful for identifications of the time of writing notes by Hilbert. Especially, the image-based search was useful.

Kazu Nagai, a historian on modern Japanese politics, happened to learn the system and applied it to his project transcribing Baron Yuzaburo Kuratomi's diary, which is said one of the most important historical materials in modern Japanese politics studies. Since the diary is gigantic, consisting of over 230 books, it is a team effort. After Nagai's ideas, some features for collaborative transcription by expert historians were added to the system. Last year, the publication of the transcription began.

Nagai projected the images of original diary and transcription on SMART-GS
system onto a screen and discussed on them with his colleagues and students. Inspired by it, Hayashi got an idea of collaborative transcription by expert historians. The style of collaboration is different from the one of Transcribe Bentham Project so that the participants get together and discuss in real time. He applied the new technique to his study of history of Japanese philosophy and he and his team are successfully deciphering documents by a philosopher Hajime Tanabe, which have been unreadable nearly 50 years.

We will analyze how and why the system were useful for the three historical studies mentioned above.

The system, which runs on Java 1.6 or later, is now distributed through sourceforge.jp under GPL2.0 license. We are now building an online platform HCP "Humanity Cyber Platform" with version control of transcriptions and annotations, and a new version of SMART-GS, by which historians can have imaged documents of online archives markup and link them.

<Keywords: transcription of imaged documents>

Tomohiko Matsumori (Doshisha University)

The objective of this study is to show the significant potential of geographical clustering using Delaunay triangulation in order to interpret the distribution of archaeological sites. The method is based on an objective procedure and the source code is distributed under the GNU General Public License and anyone can use it with R freely (http://tmats.net/pine3/).

In archaeological research, the interpretation of site distributions is an important issue. However, archeology aims at targeting the past and so it is difficult to interpret the distribution because archaeological sites give us evidence of past activities that remains only partially. In archaeology, a basic assumption is that a social relationship between neighboring sites is stronger than that of distant sites. To discover high-density areas of such sites, we would like to introduce a new objective and quantitative method.

Geographical clustering of archaeological sites using Delaunay triangulation is a method which considers the distance between the sites. Delaunay triangulation is often used in Geographic Information System (GIS) software to generate Triangulated Irregular Network (TIN) models for the representation of physical land surface. Delaunay triangulation is a set of lines connecting each point to its closest neighbors. This model makes it unnecessary to have such lines cross. This characteristic is also useful to comprehend social relationships between the sites. It is possible that short and long segments made with Delaunay triangulation reflect the strength of social relationships among the sites. In other words, short segments mean strong relationships between the sites. On the other hand, long segments indicate weak ones. By collecting the short segments from the network, it is possible to extract clusters of archaeological sites based on strong social relationships. When extracting information from the distribution of the points, many clusters are better than one. One cluster which connected all the points (sites) provides little information for archaeological interpretation. The higher the number of clusters, the more information one can obtain. In order to maximize the amount of information from the distribution of sites, it is necessary to maximize the number of clusters.

R is a free software for statistical analysis and graphics. Authors implement the procedure in R in the steps below: 1) Import latitude and longitude data of the sites into R. 2) Create a Delaunay network from the data. 3) Extract clusters to maximize the number of them from the network. 4) Export the result in Keyhole Markup Language (KML) format. The code runs without any GIS software, and the result is readable with free GIS software such as Google Earth and Quantum GIS.

The authors conducted a case study using the method described above targeting
Yayoi period sites in Toyama prefecture, Japan. We extracted 11 clusters from 60 sites of the Middle Yayoi period, 26 clusters from 150 sites of the Late Yayoi period, and 22 clusters from 121 sites of the Final Yayoi period. The result indicated that from the Middle to Late Yayoi period, the clusters were spreading. And from the Late to Final Yayoi period, some clusters disappeared while others survived.

I would like to emphasize that geographical clustering using Delaunay triangulation is useful because it provides an objective procedure for interpretation of archaeological site distributions as a basic and essential issue. Providing a simple, scientific and open method contributes to the cumulative progress of Historiography.

<Keywords: archaeology, delaunay triangulation, GIS, point pattern analysis, R>
[B-2] The spatial analysis about the Gun-ga distribution in the Ancient Japanese History

Yoichi Seino (Kyoto University)

This paper considers the strategy of local administration in the Japanese Ancient State (We call it the "Ritsuryo State") by referring the geographical information and using the FLOSS (Free/Libre and Open Source Software) GIS (Geographic Information System) called "FOSS4G" (Free and Open Source Software for Geospatial).

The Ritsuryo State set up the local government office (We call them "Gun-ga") one by one at each county (We call it "Gun") all of their controlled area. Some historians or some historical archaeologists said that the setting place of the Gun-ga was selected as the best place of the transportation or administration in the county. However they did not verify their opinions by using quantitative method up to the present time but subjective method. This is one of the serious problem in the Japanese historical scholars’ environment. Therefore, I try to show a case study. It is better way to understand the numerical method or digital information in the historical study. I used the geographical information because it is the most familiar and useful information for the historians and historical archaeologist. I try to use the GIS and DEM (Digital Elevation Model) for searching the place of the Gun-ga was actually selected as the best place inside their county. Does “Best” mean only the center of the Gun? I assume that the Ritsuryo State set up the Gun-ga at the most transportable point in the Gun. If it was so, administrators could move easily anywhere in the Gun rather than other place. So I run the Least-Cost Pathways Analysis between the Gun-ga sites, and estimate the moving potential of each Gun-ga's place. Previously, there was the study about the distributing of Koku-fu (the office of Kuni. Kuni is the upper administrative level of Gun; Province.), but authors used and show only the Thiessen polygons analysis. They did not consider the real three dimensional space but only two dimensional (the Euclidean) plane. They only showed the similarity between the Thiessen polygons and the real Province border outlines. My study can show the past human cognitive area different from the administrative area set by the state. Some of the Gun areas fits the result this analysis, but some of them does not fit it. It depended on the local circumstances at that time, the one was the relationship between neighbor Gun or transportation route, the other was the spread of the inhabitable land or cultivable land. So this is the future tasks to reveal this cause by referring more information not only the geographical data but also the historical or archaeological data.

This work is also one of the attempt to integrate some academic fields separately one by one so far, the Japanese Historical Archaeology, the Japanese History or the Japanese Historical Geography by linking Geo-spatial data or using the GIS. It tried to introduce the digital techniques, quantitative and objective methods and linking each other academic fields.

On the other hand, FOSS4G has the powerful analysis capability recently and has
many advantages for the students or freelance researchers. Keeping transparency and traceability of the algorithm is not only important in the scientific study but also in the education. Furthermore, almost FLOSS support the standard format and interoperability, so it will avoid the Vendor lock-in and keep the sustainability of data.

<Keywords: Archaeology, Japanese Ancient History, GIS, Least-Cost Pathways Analysis, FLOSS (Free/Libre Open Source Software)>
[B-2] Sharing Genealogical Spaces for Cultural and Social Anthropological Studies with the Alliance, a Kinship Database and Genealogy Management System

Shigenobu Sugito (Sugiyama Jogakuen University)

This presentation aims at pointing out the possibilities of "kinship ontology" for the computer graphics for genealogy and genealogical database management system. Professor Sugito, a member of the database development team, has a long experience in this subject as he has been doing fieldwork research in Arnhem Land, Northern Territory, Australia, since 1984. In fact the Alliance Project (which is a kinship database and genealogy management system) originated from his studies of Aboriginal society. The New Alliance project for three years started in 2010 is based on the former stages of these projects, and this paper presents the new concept of this database system, funded by JSPS. New Alliance presentation will present the followings points. Firstly, it will be useful to examine different points of view concerning the management of openness and confidentiality in databases, taking into account socio-cultural diversity, thus the New Alliance project will search for new possibilities with the help of advanced technological support. Secondly, we stress the importance of joint studies to establish database links and database management for genealogy. The Alliance Project has some experience of joint projects, as mentioned above. Flexible structures and conversion systems will be required in order to undertake joint studies with database systems. Finally, we discuss on "kinship ontology", which is a key point of the computer graphics for genealogy, and also, will contribute to the kinship study for cultural and social anthropology.

<Keywords: kinship, genealogy, database, ontology, computer graphics>
[A-3] With the time-zone differences and the busy schedules, setting up a cross-Atlantic conference call takes some time: Benefits, challenges, and strategies to support international research teams

*Lynne Siemens (University of Victoria)*

Digital Humanities is an example of a community of practice that is embracing collaboration and becoming increasingly international in focus (Siemens, 2009). New national associations in Australia and Japan have hosted their first conferences, drawing individuals from around the world (aDH, 2011; Japanese Association for Digital Humanities, 2011). Further, the Digital Humanities Summer Institute at the University of Victoria has had participants from every continent, except the Antarctic (Meloni, 2010). Finally, several large scale consortiums, such as the Text Encoding Initiative Consortium, DARIAH, CLARIN and centreNet, have members spanning national borders (centerNet, nd; CLARIN, nd; DARIAH, nd; TEI-C, 2010). The end result is potential for collaborations with team members spanning not only disciplinary boundaries, but also geographical, language and cultural ones.

To ensure that they are able to achieve their research objectives, these types of teams must find methods and means to maximize the benefits associated generally with collaboration while minimizing challenges and to address those more specifically associated with geographical, linguistic and cultural diversity (Kraut, et al., 1987; Setlock, et al., 2004; Shore & Cross, 2005). However, little is known about the best ways to prepare individuals to work in these teams and support the collaborations themselves. This paper will contribute to this discussion by reporting on the findings from diary studies which explore the experiences of individuals as they collaborate in research teams with membership drawn from a variety of countries, cultures and language groups in order to identify their benefits, challenges and strategies used to support these collaborations.

At the time of writing this proposal, the diary studies are being conducted and data analysis completed.

The benefits to the Digital Humanities community will be several. First, the study contributes to an explicit description of the community’s work patterns and relationships, particularly as the Digital Humanities community continues to become international in focus. This research demonstrates that much of digital humanities research is accomplished within interdisciplinary research teams, which are developing tools and processes to facilitate this collaboration. One particular issue highlighted in this research relates to challenges experienced within teams with members from various countries and cultures (Siemens, et al., 2009). Second, it is designed to enable those who work in such teams to recognise factors that tend to predispose them to success, and perhaps more importantly, to avoid those that may lead to problematic interactions, and thus make the project less successful than it might otherwise have been.

*Keywords:* digital humanities, collaboration, international research teams, project management>
[A-3] The Importance of Storytelling in a Digital World: Or, What the Digital Humanities can learn from TED

*Ryan Hunt (University of York)*

By conventional wisdom the Technology Entertainment and Design (TED) talks should not be popular. These talks are videos of academics and intellectuals giving powerpoint presentations for 20 to 30 minutes on topics conventionally considered too dense or uninteresting for a broader public. And yet, the TED talks are a viral sensation. As of June 2011 the TED videos had been viewed more than 500 million times.

The TED talks are a cultural phenomenon, touching on a number of issues of critical importance to the digital humanities. This paper will focus on three areas in which the TED talks excel: access, authority, and engagement. In each area TED acts as an exemplar for the digital humanities.

TED has set a high standard for allowing the internet public to access their information. TED videos can be accessed freely on a number of services ranging from Youtube to iTunes to Netflix. TED has demonstrated a keen understanding of how and where the internet public consumes content.

The TED videos also touch on issues of authority in the digital world. As the internet democratizes the access to information and allows the internet public to create as well as consume information, the authority of the ‘expert’ has eroded. The internet has granted more agency to the user, allowing anyone to be a content creator. The TED talks have embraced the user-driven web, creating a TED community which takes an active role in the talks. TED encourages users to communicate and collaborate with presenters, giving users digital platforms to engage in ideas, questions, debates.

The final, and in my opinion most important, lesson to be learned from TED is engagement. TED places a strong emphasis on the oldest of human practices: storytelling. I believe that storytelling is critical for public engagement on the web. Storytelling is a fundamentally human and social practice that allows individuals to connect through mutual cooperation and shared empathy. Storytelling inspires. Storytelling moves. It is a timeless practice that is the future for public engagement on the web.

The theme of the 2012 Japanese Association for Digital Humanities is “inheriting humanities.” Storytelling is our most valuable inheritance from the past. By combining the oldest of human practices with digital methods of collaboration and communication, the digital humanities can learn from TED’s success to create stronger ties with the digital public.

<Keywords: Social media, Web 2.0, Digital Literacy, Collaboration>
[A-3] Constructing 3D Metaverse for Inheriting and Sharing Japanese Traditional Culture

Mitsuyuki Inaba (Ritsumeikan University)

This paper presents the outcomes of ongoing research efforts on implementing 3D Metaverse for preserving and sharing Japanese traditional culture. 3D Metaverse is a platform for constructing immersive virtual spaces with digitized objects and architectures on the Internet. An avatar, or a controllable virtual character in 3D space, enables the visitors to the 3D space to participate in embodied social interactions with other avatars. These features of the Metaverse are beneficial in implementing a platform for visitors to inherit and share traditional tangible and intangible cultural heritage by situating them in social interaction, which has been difficult to realize in the conventional web-based e-Learning environments.

First, we introduce the results of our questionnaire survey to international students in Kyoto about the topics they are interested in learning by the virtual space in terms of Japanese traditional culture. Although their interest differs by gender, both male and female students showed interest in traditional architectures, gardens, costumes, and entertainments. They also showed their interest in studying traditional customs such as visiting shrines and temples.

Second, we demonstrate our Metaverse environment in SecondLife, which is the most popular Metaverse infrastructure. The environment has been constructed based on the results of the aforementioned questionnaire survey. The space includes Japanese tangible cultural properties such as Shinto shrine, Buddhism temple, and Noh stage. It also has virtual museums for Kimono costume and Yuzen textile design. Various intangible cultural heritages are also digitized and preserved in the space. For example, the visitors can perform a Noh play as avatars, whose body movements are captured from a real Noh player. They can also experience practicing manners of Shinto rituals by visiting the shrine as avatars.

Third, we explain the results of our experiments on collaborative learning using the Metaverse environment. In the experiments, both international and Japanese students interacted in the Metaverse, and exchanged their questions, impressions, and interpretations on traditional culture and customs in Japan. We analyzed their interactions and learning processes both in real and virtual worlds quantitatively (i.e., descriptive statistics) and qualitatively (i.e., thematic analysis).

Finally, we discuss advantages and limitations in inheriting and sharing traditional cultures in the Metaverse environment from a socio-cultural perspective.

<Keywords: Metaverse, e-Learning, Situated Learning, Collaborative Learning, Japanese Culture>
[B-3] DH and Disaster Management: An Overview of the UC CEISMIC Digital Archive

James Smithies (University of Canterbury)

The Canterbury region, in the South Island of New Zealand, experienced two major earthquakes during 2010 and 2011. On September 4, 2010 a magnitude 7.1 quake struck at 4.35 am, causing widespread damage and two serious injuries. Significant aftershock sequences followed. On February 22, 2011 a 6.3 magnitude quake hit at 12.51 pm. This earthquake caused severe damage and resulted in the loss of 185 lives, making it the second worst natural disaster in New Zealand history. Like the first, the second quake has been followed by thousands of aftershocks, including two significant earthquakes on June 13th and three on December 23rd, 2011.

The University of Canterbury CEISMIC Canterbury Earthquake Digital Archive draws on the example of the Centre for History and New Media’s (CHNM) September 11 Archive, which was used to collect digital artefacts after the bombing of the World Trade Centre buildings in 2001, but has gone significantly further than this project in its development as a federated digital archive. The new University of Canterbury Digital Humanities Programme – initiated to build the archive – has gathered together a Consortium of major national organizations to contribute content to a federated archive based on principles of openness and collaboration derived directly from the international digital humanities community. Two primary archive ‘nodes’ have been built by the Ministry of Culture and Heritage (‘QuakeStories’) and the University of Canterbury (‘QuakeStudies’) to collect content from the public and researchers respectively, and a ‘front window’ (www.ceismic.org.nz) has been provided by the University of Canterbury to bond the Consortium, raise funds, and provide a platform for aggregated search functions, which are powered by New Zealand’s bespoke cultural heritage schema maintained by Digital NZ. Other nodes in the federation include The Museum of New Zealand Te Papa Tongarewa, the National Library, Christchurch City Libraries, NZ On Screen, the Canterbury Museum, and the New Zealand Film Archive. The aim is to create a permanent record of digital objects for both present and future generations. To this end, the technical requirements for QuakeStudies have been reviewed by the National Digital Heritage Archive with a view to ingesting significant subsets of content (if not creating a complete dark archive) for long-term preservation. Significant attention has been paid during the design process to multi-cultural and multi-lingual requirements, to ensure content from a broad range of New Zealand communities can be ingested and researched. Future development aims to create a bi-lingual interface in English and Māori.

UC CEISMIC differs from other disaster-archives in its use of federation in addition to the development of a research-focused digital object repository to ensure broad coverage. This paper will present an overview of the project, describing the architectural model used, and the role of the digital humanities in disaster management in the
Christchurch context. Emphasis will be placed on the description of an open source model that could be deployed either before or immediately after natural disasters to safeguard the cultural heritage of affected communities, and the relationship of UC CEISMIC to other disaster archives.

<Keywords: Archives, preservation, disaster management, earthquakes, community>
[B-3] Publishing manuscript catalogues online, an attempt by the Nepalese-German Manuscript Cataloguing Project

Kengo Harimoto (University of Hamburg)

This paper documents the background, development, and future plans of the utilization of information technologies by the Nepalese-German Manuscript Cataloguing project. It is intended to illustrate an instance of a long standing humanities project adopting ever-evolving technologies, especially in the sphere of collection and distribution of information.

Nepal holds an important position in the research of indological and buddhological studies. It preserves manuscripts of texts that are unavailable from anywhere else. Particularly notable in the collection are the manuscripts of the Sanskrit Buddhist texts. The study of Sanskrit Buddhist works would have been virtually impossible, were there not for Nepal. The Nepalese-German Manuscript Cataloguing Project (NGMCP), funded by the DFG (the German Research Council) and based in the University of Hamburg, is the successor of the Nepal-German Manuscript Preservation Project (NGMPP) that ran through 1970 to 2002. The NGMPP microfilmed about 160,000 manuscripts preserved in Nepal by various institutions and private owners. The NGMCP follows up the NGMPP, by producing a descriptive catalogue of the microfilmed manuscripts.

As a background, attempts of using computer technologies prior to 2006 are discussed. The discussion includes problems present in those attempts. They may be summarized in two points: technical expertise, or lack thereof, among the project members; and the unfamiliarity with the matter in hand—in this case, expertise in indological and codicological matters pertaining to the Nepalese manuscripts—among those who were asked to provide technical solutions to the project.

The second part of the paper discusses the two online applications that the NGMCP has made available since 2006. One is an online title list and the other a catalogue wiki. The title list supersedes an electric title list first published as an MS Access application on a CD-ROM. The online title list was developed as a Ruby-on-Rails application and went online in 2007. The other online resource, the catalogue wiki is the descriptive catalogue of the Nepalese manuscripts that the NGMCP is expected to produce. It is a MediaWiki application. It went online in 2010. The rationale for choosing those specific applications, their design goals and some technical details will be discussed. The paper then refers to the observed benefits for availing the two online applications.

The third part of the paper is dedicated to discussions on the short- and long-term development plans of the applications. It talks about their integration to provide one interface to the information that the NGMCP has collected and is collecting. The possibility of delivering digital images of the manuscripts themselves will be touched upon.
The final portion of the paper discusses the issues observed during the development and deployment of online applications. Some are probably relevant to many fields of humanities. Others are unique to the NGMCP whose essential activity is to collect and publish information by many contributors, who work in two geographically separated centers; of the two, one is located in an area lacking in optimal infrastructures.

<Keywords: NGMCP, Sanskrit, Nepal, Manuscript, Catalogue>
[B-3] Australian Indigenous Languages Database, AUSTRALANG

*Kazuko Obata (Australian Institute of Aboriginal and Torres Strait Islander Studies)*

This presentation will demonstrate how AUSTRALANG (http://austlang.aiatsis.gov.au), a database of Australian Indigenous languages, serves as a resource discovery tool for Australian languages and also as a mapping tool for information on Australian languages.

Indigenous languages in Australia are traditionally not written languages and it was only when colonists arrived that the documentation of Australia’s indigenous languages began. In this short history of the documentation of Australian languages, one and the same language has been documented under a number of different names or different spellings of names, often 20 or more of them. There is no extensive standard list of names for Australian Indigenous languages (a few available ISO lists have a limited coverage) and it is often the case that resources on the same language are catalogued under different names or spellings at different collecting institutions. This situation makes it difficult for researchers to discover language resources held at different institutions, as they need to know the name or spelling used in each catalogue. A federated database of collection catalogues such as Trove or WorldCat does not offer any solution to such a situation. Users will still have to try different names and spellings of the same language to discover all relevant resources. A database of Australian Indigenous languages compiled by AIATSIS and developed by the Australian National University, AUSTRALANG, holds an extensive list of names and spellings for each language. AIATSIS will be modifying the database and upgrading the data held in the database to add a functionality to search a database such as Trove using a combination of variety of spellings and names, allowing one search to discover all relevant resources held at different collecting institutions.

The database also contains other types of information about languages, such as speaker numbers, endangerment level, documentation level, language classification, etc. It also has a Google Map interface that maps each language, and this allows mapping of languages on the basis of query results from the database. The mapping interface can be linked to other databases and map query results from other databases, promising a lot of potential as a research tool.

<Keywords: database, languages, resource, discovery, mapping>
[A-4] The *Orationes* Project: Bringing a Restoration Manuscript Online

Ilkka Juuso, Tuomo Toljamo, Lisa Lena Opas-Hänninen, Tapio Seppänen and Anthony Johnson (Oulu University)

The *Orationes* Project is an interdisciplinary initiative intended to bring an important unpublished Early Modern manuscript into the scholarly arena. The aim is to create a digital archive which will make the texts available to a wider audience. The manuscript, preserved as Lit. MS E41 in the archive of Canterbury Cathedral, was collected, and in part composed, by George Lovejoy (c. 1675), Headmaster of the King’s School, Canterbury, after the English Civil War. The texts within it represent one of the most substantial unpublished sources of English School Drama from the period. As well as containing a previously unnoticed adaptation of a pre-war play by a major author (James Shirley), this large volume, comprising 656 folio pages and running to some 230,000 words, includes a number of short plays and dramatized orations written in English, Latin and Greek by the scholars and staff of the King’s School. These works celebrate the Restoration of Charles II to power, re-enact the Gunpowder plot, and discuss a wide range of topical issues.

The aim is to create a rich digital edition that utilizes the high resolution scanned images of the manuscript pages and the TEI-compliant transcriptions created by domain experts. We are approaching the work from two opposite directions: firstly, we are creating an uncompromising TEI-XML transcription of the manuscript, which not only entails rigorous manuscript transcription but also the translation of the Latin and Greek portions of the material and the identification of features of interest for use in a rich visual interface. Secondly, we aim to create a reusable software package that does not sacrifice functionality or source integrity, but can nonetheless be used to create digital editions from similar XML and image source materials in a straightforward manner. Combining these two goals successfully requires navigating the TEI guidelines, creating an automatic linking mechanism between image and text, creating efficient search mechanisms for TEI data and creating an interface that is both intuitive and generic at the same time. It is these problems that we would like to present, discussing some possible solutions and inviting conference participants to suggest other solutions.

*Keywords:* digital archive, visual interface, encoding, linking image and text, Restoration school drama

Shoichiro Hara (Kyoto University) and Motomu Naito (Knowledge Synergy Inc.)

MANGA is a picture-oriented visual media with written-voices to convey narrative or information and structured by its own visual grammar (e.g., frames, balloons). There many opinions on the difference of MANGA and comics, but this abstract treats MANGA to include comics.

MANGA is said to establish itself a contemporary culture. Its market expands worldwide and it is distributed by various media such as papers, Web and mobile phones. Metadata for the new cultural media must be necessary to organize its complex structure and to give access-methods to consumers under the web environment.

A distinguishing feature of MANGA is its media-diversity. We can observe it from the two points of views.

The first is its complex derivation. Most of MANGA is originally published as weekly magazine articles. And later, publishers compile books from articles. Some books are turned into media conversion such as novels, dramas and movies. Some books are exported to overseas countries and then translated and turned into media conversion. The second is its unique structure. MANGA is composed of pictures, texts and symbols, but different from picture books, they are tightly coupled in accordance with its peculiar grammar and semantics. MANGA is also different from a movie, but a sequence of pictures can express dynamics and/or time sequence of stories like a movie.

This paper will describe our challenge of constructing MANGA metadata to organize media-diversity and of developing tools to manipulate metadata.

1. Bibliographic Metadata:

Bibliographic metadata describes elementary information of each content such as an article, a book, a novel, a movie etc. Some research papers have already proposed MANGA bibliographic metadata, but as they are designed to describe much information (e.g., derivation and media conversion), they become very complex; nevertheless, they are not enough to describe all information.

We consider that metadata should be simple and constructed of de facto standard element sets easily to create and manipulate them. And as for MANGA, rich information about MANGA work (e.g., an author, a publisher, bibliographical notes etc.) has already described and distributed on the Web, we will use the information. That is, our bibliographic metadata is designed to describe information of a single content, which is based on Dublin Core Metadata and is used as a container to store URL to detail information on the Web. Relationships among contents are described by another framework (Topic Maps).
2. Structure Metadata

Basic structure of MANGA is hierarchical (i.e., work is composed of books, a book is composed of pages, a page is composed of frames, and a frame is composed of pictures, texts and symbols), and our structure metadata follows this structure. In MANGA, boundaries between frames are sometimes ambiguous, which make difficult to define the sequence of frames. In these cases, texts are read carefully to arrange frames to create a smooth story. As translated works have the same structure of the original ones, creation of links between translated works and original ones are simple.

3. Scene Metadata

A scene is a story of MANGA composed of a sequence of frames, which is a similar way of a movie and a TV drama. We make scene metadata by following procedure:

- Finding a changing point of stories and looking on it as the end point of the previous scene and the start point of the next scene. Location, time and personage in pages should be strong clues to find points though, human has to decide it in the present.
- Creation of metadata of each recognized scene. Subjects and title of the scene are given in addition to the location, time and personage as the items of the metadata.

4. Topic Maps for Metadata Network

In our method, overall MANGA metadata is composed of simple metadata, that is, it begins from top level bibliographic metadata that describes general information about work to scene metadata that describes a story. Relationships among metadata (i.e., derivative versions) form a network of metadata. We introduce Topic Maps to describe the metadata network. The tolog, a topic map query language, is available to search for a specific book, a scene, a text and so on.

<Kwera: MANGA, metadata, Topic Maps>

Kazushi Ohya (Tsurumi University)

From 2008 to 2010 we had undertaken the project of language documentation for endangered languages in North-East Eurasia. In the project, we made an experimental system, which handles corpora made by a software ToolBox and which validates and transforms them into two kinds of XML data: one based on an original simple scheme and the other based on a scheme in a standoff style. The system is implemented by using Berkeley DB XML as the database engine and XQuery as the query language, and provides a retrieval web service (OHYA 2011).

Requirements of descriptive linguistics are summarized as follows:
1. sharing language data,
2. keeping an interlinear format on every stage of data processing,
3. handling multiple data units and their structures,
4. using many characters including IPA in entering, editing, and retrieving,
5. expecting a software system to support entering or encoding language resources in the same way in each process on the PC,
6. selecting a scheme from multiple choices.

The present difficulties in language documentation especially for a study of endangered languages are summarized as follows:
1. defining a digital platform to record and handle language data,
2. transforming language data into a specific form expected in global-scale repositories or in printing.

In our observations, these difficulties seem to relate to the gaps between requirements from linguists engaged in descriptive linguistics and means provided by researchers in CS especially NLP.

Linguists have been seeking a way of consistent data sharing in language documentation applied in both a personal environment and a research and language community. They can be called personal diachronic sharing and community synchronic sharing respectively. On the other hand, researchers in NLP have been seeking a way of community synchronic data sharing only: collecting language data, automatic parsing and tagging, and finding a plausible candidate or rule from the whole data. Global-scale repositories led by researchers in NLP require linguists to convert their own data into valid data to the system without data conversion services.

In order to realize personal diachronic and community synchronic data sharing, there have been three ways:
1. using a common platform/application,
2. sharing a common scheme,
3. sharing a common model.

However, from our previous project, they do not function well. For example, as for common scheme, LAF/GrAF proposed by ISO contains many problems:
(1) lack of link path management,
(2) lack of sharable query format on such multi-link-path model,
(3) minimum unit constraint on primary data,
(4) multiple types of primary data that prevent data sharing,
(5) both primary and unit definition data constitute a pair of data units that detracts reusability,
(6) lack of general wrapper of multiple marked-up data with tree-structured model.

In the new project, we will adopt a new strategy of using a common API for sharing data. In this presentation, we explain in detail about our previous and present projects.

<Keywords: language documentation, markup language, data sharing>
[B-4] Panel: Undergraduate Models for Inheriting Interdisciplinary and 
Humanistic Studies in the Digital Age

Alex Benkhart, Janet Thomas Simons, Kyoko Omori and Angel David Nieves 
(Hamilton College)

As Giles Gunn has argued, interdisciplinary studies “can be mapped only when one 
redraws the critical coordinates supplied by those disciplines” through applications that 
transform or transgress traditional methodologies and approaches. This panel presentation 
will focus on the role of undergraduate institutions in advancing digital humanities as a 
growing field of study among small, elite liberal arts colleges. The Digital Humanities 
Initiative (DHi) at Hamilton College is different from existing digital humanities centers in 
that it models interdisciplinary collaboration across academic and administrative units 
while prioritizing curricular integration and undergraduate collaborative research. These 
activities support a fundamental shift in humanities-based research, leveraging the 
potential of technology to access and manipulate rich media collections in ways that 
increase collaborative interdisciplinary scholarship and lead to the generation of new 
knowledge formation. Liberal arts colleges are uniquely poised to address the challenges 
facing digital humanities, in particular, the development of sustainable infrastructure 
architecture for faculty research projects and interdisciplinary models for collaborative 
scholarship. Benkhart, Simons, Omori and Nieves each detail how their projects brought 
together practitioners and researchers of information technology with digital humanists to 
break new ground in their respective fields including Asian Studies, instructional 
technology, Japanese Studies, and Africana Studies.

Benkhart Presentation Abstract:

Benkhart examines how homosexuality is depicted in popular culture, particularly in 
movies and anime. This examination is strongly geared towards sociolinguistic analysis,
incorporating Makoto Furukawa’s previous work on “codes.” Benkhart proposes that in 
popular culture, sexuality is often seen and described in terms of asobi and shumi, play 
and hobby/taste, respectively, especially in the cases of homosexuality and 
transgenderism. This work briefly analyzes Japanese movies and anime with LGBTQ 
themes made over the past 20 years with a focus on Yoshihiko Matsui’s “Where Are We 
Going?” and Nagisa Oshima’s “Taboo.” Benkhart started much of this research while 
working at Hamilton’s Digital Humanities Initiative on Prof. Omori’s Japanese 
Comparative Film Archive Project.

<Keywords: queer, codes, anime >
**Simons Presentation Abstract:**

Simons presents collaborative methods and processes at Hamilton College’s Digital Humanities Initiative (DHi) that have led to the "creation and curation of humanities digital resources” and have integrated digital humanities into “academic curricula." DHi at Hamilton College is a collaborative where new media and computing technologies are used to promote humanities-based teaching, research, and scholarship across the liberal arts, www.dhinitiative.org/.

*<Keywords: sustainability, digital infrastructure>*

**Omori Presentation Abstract:**

Omori presents on her DH project currently developed at Hamilton College for interdisciplinary analyses of silent films (images) and Japanese “benshi” performance (audio). In their heyday during the 1920s, thousands of benshi performers wrote individual scripts for the same movies and performed alongside a screen as a live narrator/commentator/voice actor. To study the ways in which benshi’s voice created audio-visually complex narratives, this interface analyzes benshi’s scripts, translates them, syncs movies and their benshi audio clips by different benshi, and study the interplay between the image and audio. Omori and Hamilton DHi are also seeking ways to digitally recreate the performing space of a movie theater in early twentieth century Japan as part of this project. In the process of developing the interface, Omori has trained three Hamilton students on the study of Japanese cultural production, especially silent film culture.

*<Keywords: silent film, Japanese benshi performance>*

**Nieves Presentation Abstract:**

The Soweto Historical GIS Project (SHGIS) seeks to build a multi-layered historical geographic information system database that explores the social, economic and political dimensions of urban development under South African apartheid regimes (1904-1994) in Johannesburg’s all-black township of Soweto. The application of geographic methodologies to the study of the anti-apartheid movement reveals the complex spatial dimensions of violence, resistance, and freedom. The project examines the micro-geography of resistance and the layering of meaning and action between the apartheid state and township residents across its built form. By documenting across space and time the racial and political ideologies of apartheid within these townships, or Black labor-machines, an important question is raised: can we map residents’ resistance to oppression using GIS? Research suggests that modernism – as expressed through urban planning and architectural design – was upended by the ground-up activism of township residents in the struggle against apartheid. Nieves has trained four undergraduate research assistants from the Department of Geography, specializing in historical GIS, at Middlebury College.

*<Keywords: historical GIS, apartheid>*
[C-1] Quantitative analysis of style change and conversational sentence within the works of contemporary novel writer

Akira Kudo (Tokyo Institute of Technology)

The overarching goals of our research are to scientifically interpret literature, as a form of art that closely reflects higher human sensibilities, and to verify the stylistic changes within one writer of novels. In seeking to achieve these goals, the aim of this study is to quantitatively analyze texts using some techniques from the digital humanities.

Although traditional methods of literary study offer various interpretations, it is difficult to assert the consistency and objectivity of such methods because some degree of arbitrariness and researcher subjectivity is inevitable when handling research materials. Accordingly, this study adopts a stylo-statistical approach that has been developed within the field of bibliometrics, in order to ensure consistency and objectivity, as a method for literary studies.

The target data consists of the twelve full-length works of Haruki Murakami, who has been a representative of modern Japanese literature since the 1980s in developing a unique style of story and writing. Murakami has received not only a domestic literary prize but also high commendations from overseas, such as winning the Franz Kafka Prize, the Frank O'Connor International Short Story Award, and the Jerusalem Prize. As a living Japanese writer, he is exceptionally rare in possessing a form of universal literariness that is recognized all over the world. Generally speaking, many researchers have commented that Murakami's style has undergone a shift during the 1990s, a period that witnessed the Aum Incident. However, the gradual stylistic transformations across his twelve works have not been discussed from an historical perspective.

We have made electric texts of the twelve full-length works using optical-character-recognition (OCR) technology. This study creates word frequency lists to extract feature vectors that classify the vocabulary in terms of both word-class and semantic categories, and conducts clustering analyses on the data in order to identify groupings based on style within these works. Specifically, we have used the National Institute for Japanese Language’s (2004) "Word List by Semantic Principles, Revised and Enlarged Edition", which is a Japanese dictionary of synonyms as a means of grouping the words into semantic categories.

The analysis based on word-class yields clusters that reflect diachronic divisions that indicate that Murakami’s style has shifted over time. In contrast, the analysis based on semantic categories yields clusters that reflect the themes and the contents of the novels, indicating that Murakami’s interests have also shifted, moving from the individual to society.

In addition, this study analyzes the descriptive and conversational sentence within Murakami’s "1Q84" using measurements for sentences, words, letters, and vocabulary,
and conducts clustering analysis on the character conversations in order to identify groupings based on conversational style of each “1Q84” character. The results yield clusters that represent three categories reflecting the three character functions of subject, helper and sender. We also look at these categories which resemble Greimas’ actantial model which claims that the story has six types of character.

Finally, through a systematic comparison of our quantitative method to more traditional methodologies of literary studies, we discuss the effectiveness of the approach for literary interpretation.

<Keywords: quantitative analysis, novel, style change, conversational sentence, Haruki Murakami>
[C-1] Logical Layout Analysis for Japanese Journal in Humanities with Machine Learning

Katsuya Masuda and Hideki Mima (University of Tokyo)

The purpose of the study is recognition of logical types of blocks in OCR results using machine learning techniques, without constructing rules for the recognition by hand. The approach using machine learning techniques can be applied to any types of layout structure, if these are adequate amount of correct documents for the type of layout structure.

Large amount of digitalized images for historical documents have been produced with progress of document digitalization and digital libraries. Extraction of text in the images is important especially for historical documents, for which digitalized text data do not exist yet, and useful for supporting systems for accessing documents. A search system over a large amount of historical documents and analysis of the documents, such as information extraction and ontology construction for the concepts in the research area of humanities, will be very helpful for researchers in humanities.

Although recognition of characters, blocks of texts and figures in document images has been improved by the development of the optical character recognition (OCR) techniques, the identification of the logical structure of recognized blocks, logical layout analysis, are still a difficult problem. By the logical layout analysis, the logical types of blocks (e.g. title, author, caption, header in the documents) are identified. Detection of logical structure is indispensable for accessing or analyzing documents. For search systems, bibliographic information such as titles or authors of documents is key information for searching documents. For analyzing systems, bibliographic information is also important and detection of blocks other than body of documents, such as the page number, the reading header and the caption of figures or tables, is necessary because these blocks should be ignored for analyzing documents.

For the logical layout analysis, many rule-based approaches using classification rules constructed by hand[1,2] are developed, but the construction of rules requires very high cost and the rules constructed for a particular type of document layout cannot be applied to the documents in different types of layout. Some approaches using machine learning techniques[3,4], which enable automatic construction of classifiers from the correct data, have been developed in recent years, but the target of these approaches is an extraction of only bibliographic information or reference information from the papers in the scientific area. These techniques cannot be applied straightforwardly to the articles in humanities, because the layout structure of documents is different from the scientific papers, so the features used for detecting the logical type of blocks are different.

We proposed an approach to identify a logical type of the blocks using machine learning techniques. We employed the Support Vector Machine (SVM) [5] as a machine learning technique, and extract various types of features from the OCR results. We
employed the following features to identify the logical type of blocks:

**Positions**: x and y coordinate of the block

**Blank Space Length**: the length of a blank space for four directions (upper, lower, left and right)

**Block Size**: the width and height of the block

**Character Size**: average width and height of characters in the block

**Noun**: percentage of “noun” words in the text of the block

Proposed approach and existing rule-based approach were evaluated on the OCR data of one of the Japanese journal in humanities, “Shiso(Thought).”

In the experiments, the systems classify blocks into five logical types, title, author, header, page number and body. Table 1 shows precision and recall of recognition by two approaches for each logical type. The results show that proposed approach with machine learning overcomes the rule-based approach, especially in the identification of “TITLE” and “AUTHOR” blocks, which are important for recognizing the unit of papers.

### Table  Precision and Recall of Logical Type Recognition

<table>
<thead>
<tr>
<th></th>
<th>Proposed approach</th>
<th>Rule-based approach</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>TITLE</td>
<td>0.976</td>
<td>0.945</td>
</tr>
<tr>
<td>AUTHOR</td>
<td>0.996</td>
<td>0.974</td>
</tr>
<tr>
<td>HEADER</td>
<td>0.974</td>
<td>0.984</td>
</tr>
<tr>
<td>PAGENUMBER</td>
<td>0.994</td>
<td>0.996</td>
</tr>
<tr>
<td>BODY</td>
<td>0.993</td>
<td>0.992</td>
</tr>
</tbody>
</table>

**References**

Analysis Systems, pages 117–125)

<Keywords: Layout Analysis, OCR, Machine Learning>
[C-1] Stylistic analysis of tweets that are likely to be shared

Yui Arakawa (University of Tsukuba), Akihiro Kameda (University of Tokyo), Akiko Aizawa (National Institute of Informatics) and Takafumi Suzuki (Toyo University)

This study investigates the characteristics of Japanese tweets that are likely to be shared. Twitter is a microblog that has many users throughout the world, including in Japan. One of the important purposes of using Twitter is to distribute information. For effective distribution, it is useful to know the characteristics of tweets that are likely to be shared. Two essential factors for "sharing" tweets are (a) tweets that are likely to be retweeted (RT), and (b) tweets that are posted by special types of users who have a large number of followers. Therefore, we decided to investigate the characteristics of (a) and (b). To investigate these two types of tweets, we proposed special features specific to tweets, namely, (a) appearances of special symbols (RT, @, #) and URLs, as well as (b) information roles of tweets (monologue, dialogue, advertising, informative) assigned by manual content analysis. We added these features to conventional features that are used in stylistic texts analysis, namely, (c) appearances of function words, and (d) ratio of part-of-speech tags. We gathered 28,756 tweets sent by the top 10 accounts with many followers, and applied random forests classification experiments for assigning labels of high/low RT tweets and four user categories (entertainer, celebrity, organization, and character/bot), all of whom have a large number of followers. The results showed that our proposed features were very important for these two types of classification experiments. Using our proposed features that are specific to tweets as well as conventional features used in stylistic text classification, we can determine the features that influence classification in two ways. By clarifying the characteristics of tweets that are likely to be shared, this study contributes to the effective information that is likely to be shared in Twitter. In addition, this study provides an important case study for computational stylistics, a subfield of digital humanities, by investigating the characteristics of new types of texts.

<Keywords: content analysis, Twitter, random forests, retweets, stylistic text analysis, tweets>
[C-1] On a Digital Database of Ancient Egyptian Hieratic Palaeography

Masakatsu Nagai (University of Tsukuba)

The language that was used in Pharaonic times is known as Ancient Egyptian. In order to express Ancients Egyptian in writing, for the most part, hieroglyphs and hieratic glyphs were employed. Hieroglyphs were like block-style letters, while hieratic glyphs were cursive, freely written forms. Research in the palaeography of the hieratic script was extremely late. One reason for this is that in Egyptological research, scholars recognized texts that had been transliterated from the original hieratic into hieroglyphs as primary sources. However, it is impossible for transliterated texts to be primary sources. Furthermore, even though it is now a century old, G. Möller’s Hieratische Paläographie (Leipzig, 1909–12) is still considered the definitive work on hieratic palaeography today.

This being the case, this applicant, because of the need for a basis for promoting research in hieratic paleography, is in the midst of creating a database of digital palaeography that uses photographs of actual hieratic glyphs. There are two types of data that are required for this database. The first is meta-information on the glyphs, such as the glyph code, its color, its pronunciation, etc. The second is a digital photograph of each and every glyph included. Up until the present, these two types of data were linked by creating a search system with an existing software application (FileMaker Pro 11, Ver. 11). In this system, if one searches by using the glyph code, data for the relevant glyph appears, together with photographs of the actual glyph. At that time, the actual phrase containing the glyph is also displayed, making it possible to obtain an exhaustive amount of information concerning the hieratic glyphs found in Möller.

In addition, this applicant has also made digital images of the glyphs, and created a retrieval system in which one can search for them. If possible, he would like to go on to make preparation to present a system that links his digital hieratic palaeography with Möller’s palaeography.

To the best of the applicant’s knowledge, this digital palaeography of Ancient Egyptian hieratic glyphs is the first of its kind in the world. Although the retrieval system itself is still undergoing research and development, it is believed that the very attempt to create it will have a great impact upon researchers around the world. As this presentation will also include a demonstration of the retrieval system, the applicant hopes to obtain valuable feedback and opinions from the participants.

<Keywords: Digital Palaeography, Codicology, Egyptian Language, Hieratic Manuscripts>
[C-1] Design of Serial Comparison Model for the Diachronic Corpus Study of Japanese

Hilofumi Yamamoto (Tokyo Institute of Technology / University of California, San Diego), Makiro Tanaka (National Institute of Japanese Language and Linguistics) and Yasu-Hiro Kondo (Aoyama Gakuin University / National Institute of Japanese Language and Linguistics)

The project “the design and development of a diachronic corpus of Japanese” began in 2009 at the Department of Corpus Study, the National Institute of Japanese Language and Linguistics (NINJAL) as a collaborative research project by linguists and scholars of literature from NINJAL and the University of Oxford. Its focus is on collecting representative Japanese literary works and classical documents from the tenth century to the nineteenth century.

We are currently working on the development of a prototype version of the diachronic Japanese corpus: i.e. we are focusing on the selection of materials, digitizing texts, adding alternative texts (different orthography) to original texts, compiling a basic thesaurus that differentiates between different spellings, and word segmentation.

This paper addresses the discussion of the basic concepts encountered during our work on the project: synchronic and diachronic analysis, which led us to the design of a serial comparison model which allows us to examine language change between documents or literary works with respect to time.

Saussure (1983) defines diachronic change as a language feature transforming according to time, and synchronic as a language feature at a certain moment in time. In either event it is necessary to compare any two types of texts to examine differences in terms of diachronic or synchronic change. What we examine is not differences in content but those in language use. Therefore we must extract, in a clear way, only the differences in language use between the target texts.

Many studies have tried to examine differences in language use. However, in almost every case, examples or clues of language change were collected not by computers but by time-consuming manual work. The methods they employed are thus impractical for the collection of a large number of examples or clues, as is often the case in corpus studies. Furthermore, there is a risk that utilizing such methods could lead researchers into only collecting sentences conforming with their hypotheses. Therefore, it is necessary to use computers and corpora for conducting more exact and unbiased research than was ever achieved or possible before. This, however, requires us to prepare a framework that can clearly differentiate ways of language use and contents of texts.

To this end, we developed a “serial comparison model” which allows us to examine evidence of language change by searching in corpora, and classify vocabulary items into
differences in content or differences in language use.

The finding of this study is that the serial comparison model allows us to 1) discern language change between any two points in time, and 2) classify vocabulary items into five categories: unchanged, similar but not used in the new text, similar but not used in the old text, newly used, and abandoned. When comparing two texts, the model extracts differences in language change from differences in content.

<Keywords: Corpus Linguistics, Diachronic Language Study, Comparative Study>
[C-1] Emotive Adjectives and Verbs of the Heian Japanese

Makiro Tanaka (National Institute for Japanese Language and Linguistics) and Hilofumi Yamamoto (Tokyo Institute of Technology)

In Modern Japanese, agents or objects of emotive adjectives can be indicated as follows: 'watashi wa kanashi' (I am sad) or 'wakare ga kanashii' (farewells are sad). It is also possible to use emotive verbs in the modern written style called 'gendai kakikotoba': 'wakare o kanashimu' (sorrow over farewells).

On the other hand, the presence of agents of emotive adjectives and sentences containing emotive verbs is missing in the Tale of Gengi (in the Heian period; ca. 1000; Genji Monogatari) which is written in traditional Japanese sentence style called 'wabuntai'. The characteristics of sentence structures in the Heian period are summarized as follows:

- Agents are not indicated in adjective sentences.
- The use of verb sentences depends on the kind of emotion the verb indicates.

The aforementioned grammatical usages have turned out to be different in our examination of the corpus of Tales of Times Now Past (Konjaku Monogatarishū; ca. 1130; established at the end of the Heian period) which consists of both 'wabuntai' sentences and 'kanbunkundokutai' sentences. ('kanbunkundokutai' is a style of written Japanese that follows the grammar of Chinese classics (kanbun) and reproduces the style and feel of Japanese sentence sounds (kundoku), although it was never transcribed in Chinese classics.)

The change is that we clearly observed the use of emotive adjectives and the use of various kinds of emotive verb sentences, whose sentence constructions are very similar to those of modern Japanese. By comparing sentences in Tales of Times Now Past with equivalent sentences in Chinese texts in the original works of Tales of Times Now Past, we assured that this change of sentence construction structure is caused by the effect of Chinese sentence styles.

Therefore, we estimate that the two styles (wabuntai and kanbunkundokutai) merged in the 12th century and the use of emotive agents or objects and emotive verbs, as seen in contemporary Japanese, are a direct result of the mixing Chinese sentence structure (kanbunkundokutai) in with the language.

<Keywords: corpus linguistics, diachronic language study, comparative study, stylistics of classical Japanese, mix of two styles>
[C-1] Toward Non-Consumptive Formal Evaluation Challenges Using the HathiTrust Research Center Digital Collections

J. Stephen Downie (University of Illinois Urbana-Champaign), Beth Plale (Indiana University), Robert H. McDonald (Indiana University) and Marshall Scott Poole (University of Illinois Urbana-Champaign)

The HathiTrust Research Center (HTRC) is a collaborative research center launched jointly by Indiana University and the University of Illinois to act as the public-facing research arm of the massive HathiTrust Digital Library. The HTRC is mandated to help digital humanities (DH) researchers from around the world surmount the difficulties associated with the successful processing and analysis of terascale amounts of digital text. Thus, the scholarly developers at HTRC work daily to develop cutting-edge software tools and cyberinfrastructure to enable advanced computational access to the growing digital record of human knowledge. HTRC began its efforts July 2011.

We at the HTRC propose engaging the DH community of scholarly developers in the creation and running of series of formal evaluation challenges. These challenges have two simultaneous goals. The first goal is to encourage the development of novel techniques that can be used to improve the quality and utility of the massive HTRC datasets. The second goal is to encourage the engagement of the DH community with the collections of some 10 million items being made available via the HathiTrust.

The HTRC challenges are inspired in part by the Music Information Retrieval Evaluation eXchange (MIREX)(Downie 2010). MIREX, run by HTRC co-director Downie, has evaluated some 1400 different algorithm runs over a wide range of music information retrieval (MIR) since its inception in 2005. Like HTRC, MIREX cannot directly share its data collections with researchers. To overcome this limitation, it MIREX developed an “algorithm-to-data” approach to “non-consumptive” tool evaluation. Under this model, algorithms are sent into the MIREX labs where they are securely run by MIREX technicians against the locked-down datasets. The results of evaluation and NOT the underlying data are the only items returned to the evaluation challenge participants. MIREX runs on an annual cycle with task proposed in early spring and evaluation run conducted over the summer each year. The results are returned to participants in early fall. Participants then meet at a dedicated MIREX plenary and poster session held in conjunction with the annual International Society for Music Information Retrieval (ISMIR) conference in late fall. At this meeting, participants compare results and brainstorm new challenge task ideas for the next MIREX iteration.

MIREX is generally acknowledged for having been a key factor in the successful establishment of the vibrant and thriving MIR research community. MIREX has also helped many student researchers advance their understanding of MIR tools and techniques. MIREX has been praised by more senior researchers for advancing the state-
of-the-art in MIR algorithm development: MIREX results are now considered the baseline from which new work must progress.

The key to MIREX’s success and impact is its fundamental commitment to being a community-led endeavor. It is the community, and not the MIREX organizers, that defines the evaluation challenge tasks to be run each year. It is the community that defines the datasets to be used. The community also defines the metrics by which success is judged. The net effect of this community involvement in MIREX makes MIREX an integral part of what it now means to be a member of the MIR research and development community.

In the spirit of MIREX, we are proposing to the DH community that they consider the creation of an evaluation challenge designed to improve the quality of the OCR-derived text that represents each document. Another potential challenge being put forward involves tools and techniques to substantially improve the quality and quantity of metadata associated with each document. The corrected and enhanced data would become part of the HTRC data and available to all researchers.

We also envision the DH community trying out, and then evaluating, a wide range of novel machine learning and information retrieval techniques. For example, other evaluation challenges might also deal with automated author gender identification, genre classification of materials, and major themes identification of the works in the collection. Again, as this all about building community, it will be the DH community that actually defines the final set of tasks to be evaluated.

<Keywords: Evaluation Challenges, Digital Humanities, Non-consumptive Research, Google Books, Community Building>
[C-1] Interactive display of shining gilded illumination in Keio Gutenberg bible

Hiroyo Ishikawa, Fumihiko Kamemura, Kunitake Kaneko, Hideo Saito and Takami Matsuda (Keio University)

In this poster presentation, we will introduce a system for displaying digitized rare books, which provides realistic observation by changing appearance of the virtual light direction of the rare book in interactive manner.

Recently some rare books have been provided for public reading as digital data. However, almost all pages are just photographed or scanned as raster images of each page in the books. If the purpose is to read the sentences of a rare book, showing such digital raster images of pages is sufficient for making people read the book. If, however, the purpose is to feel the realistic appearance of the book, we may wish to observe from not only one direction, but also different directions, especially when the book includes pictures, illuminated letters, such as art paintings.

The old books printed in the middle ages of Europe, called "incunabula", were illuminated by handwriting. Such illuminations were painted for initial letters and blank margins of pages. Some illuminations were gilded for making the surface shine as a gold. Such golden shining cannot easily be represented by simply showing a digitized image of the surface of the page, because the golden shining can be observed as change of brightness by viewing from different direction. Therefore we aim to develop a system that enables users to interactively observe such illuminated letters of the rare book from various angles, so that they can feel the realistic appearance of the surface of the book.

As the displayed contents, we choose digital data of the Keio Gutenberg bible, which was photographed in the late 1990’s. The digitizing of a rare book is not easy and photography or measurement is not possible many times. Therefore our purpose is to display illuminations of a rare book effectively by using past digital data (a color image) photographed from the front of a page.

For achieving such interactive display of the Gutenberg bible, we choose iPad3 that has multi touch screen and a high-quality display as a platform of the system. We aim to make users to interactively change the angle and the scale of a displayed three-dimensional (3D) model by their fingers.

The contents that are displayed in the iPad3 are also very important. For the contents, we created a virtual 3D model of the page with color textures. The 3D model is generated by observing photographed images and the bend of a page. Then the bulge of a gilded illuminated part and the printing part are added to the model. The photograph without highlight is used as a texture.

For real-time rendering in the system, some virtual light sources are set, and then
display image is generated according to the virtual observation direction. In addition, a specular reflection is produced only to a gilded part by using a specular reflection map with the virtual light resources in GPU.

As a result, we can observe interactively illuminations of the Gutenberg bible changing the direction of a page. Additionally, at the part of gilded illuminations, we can observe sparkle of the surface effectively. Future works are 3D reconstruction of a page and a whole book by using some techniques, for example photometric stereo or stereo camera.

The following photo shows the developed system.

<Keywords: interactive display, surface illumination, rare book, Keio Gutenberg bible, shader, specular rendering>
[Invited Poster]
A Quantitative Investigation of English Accents

Martijn Wieling (University of Groningen and University of Tübingen), Jelke Bloem (University of Groningen), Kaitlin Mignella (University of Groningen), Mona Timmermeister (University of Groningen), R. Harald Baayen (University of Tübingen), and John Nerbonne (University of Groningen and Freiburg Institute of Advanced Studies)

We examine a large database of English pronounced by second-language learners and we measure the degree of deviance of the pronunciations from that of standard English.

The proponents of a critical period for language learning (Penfield & Roberts 1959, Lenneberg 1967) hypothesize that there is a maturationally constrained *critical period* after which language acquisition is ultimately less successful. Ignoring studies on deaf children brought up in non-deaf households (Newport 1990), most recent work focuses on the consequences of the critical age hypothesis for second-language learning. Because children brought up without language are rare and generally at a disadvantage in many ways, studies on them are also rare and difficult to interpret.

If a critical period for language learning exists, then the age of onset for learning a second language should influence its success, and there should be breaking points, i.e., watersheds after which learning is less successful (DeKeyser 2012). Various studies have identified possible breaking points, but these vary per study (Singleton 2005, Muñoz & Singleton 2011). Here we show, based on a large collection of pronunciation data (Weinberger & Kunath, 2011; more than 800 learners) and a validated measure of pronunciation deviance (Nerbonne & Heeringa, 2009), clear support for breaking points for native speakers of both Indo-European (IE) and non-Indo-European (non-IE) languages. The addition of breaking points (in a "mixed model" regression analysis) to the model enables us to explain 1% more of the variation. The breaking points occur at very different ages, however, namely at 18 (IE) vs. 6 (non-IE), suggesting that the critical period is not a simple maturational constraint, but rather a constraint that interacts with first languages. Learning a first language interferes with learning a second (Birdsong & Molis 2001), and we conjecture that less similar languages interfere in ways that are more difficult to overcome. This explains our own data and suggests a reason for the otherwise puzzling variability in results reported in previous studies, as these studies needed to attend to the languages involved.

References


[Workshop]
Introduction to Document-based and Genetic editing in TEI

Lecturer: Elena Pierazzo (King's College London)

Significance of the workshop
In the past twenty years the Text Encoding Initiative (TEI) has established itself as the de facto standard for the representation of digital texts and the fundamental basis for each digital scholarly edition of historical documents and literary manuscripts. Yet the TEI model has been criticised by many scholars for privileging the textual, linguistic aspects of the primary sources with little awareness of the complexities of the layout and structure of the original supports other than for their description. During the past four years, the TEI’s Special Interest Group on Manuscripts (chaired by Elena Pierazzo, proponent of this workshop) has elaborated a new proposal. This revolutionary new approach allows for the encoding of texts from a documentary point of view: that is, page by page and line by line rather than (or alongside) chapter by chapter and paragraph by paragraph. This new approach also includes support for the encoding of genetic criticism and is particularly well suited to modern, draft manuscripts. This new way of encoding has become an official part of the TEI since December 2011 and represents its most substantial innovation since the introduction of the P5 version of the Guidelines in 2007.

Methodology
The workshop will introduce participants to the transcription of primary sources (draft manuscripts, correspondence, diaries) from the genetic criticism and documentary points of view. It will consist of an introduction to the new elements and of ‘hands-on’ sessions where the participants will be able to put them to use by encoding both their own material and that provided by the instructor. As the new elements include a way to encode the documentary view of manuscripts and not only genetic criticism, it is expected that the workshop will be of interest to anyone with an interest in editing manuscripts, not only to modernists.

The workshop assumes a general knowledge of XML and the TEI.
Participants are requested to bring a laptop (PC, Mac, Linux) and ideally to have their own material at hand for practicing.
Software required (preinstalled):

• oXygen by SyncRO Soft Ltd (v. 12 or later), multiplatform and available on a 30 days free licence from http://www.oxygenxml.com
• Inkscape, multiplatform and available for free from http://inkscape.org
• Web browser supporting SVG, i.e. an up to date version of Chrome multiplatform and available for free from http://www.google.co.uk/chrome or Safari v.5 and later for Mac and Windows available for free from http://www.apple.com/safari/download/

Timetable

The workshop is planned for a full day with the following preliminary timetable:

Morning session (3h30’)

• Introduction (30’)
• Documentary view (1h30’)
• Exercises: material prepared by the instructor
• New elements for transcription (1h30’)
• Exercises: material prepared by the instructor

Afternoon session (3h30’)

• Documentary encoding in real life (1h)
• Exercises: material brought by the participant
• Genetic criticism TEI-style (1h)
• Time-based encoding
• Exercises: material brought by the participant and/or provided by the instructor
• Outputting documents via SVG (1h)
• Discussion (30’)
• Problems, concerns
• Feedback
[Pre-conference Lecture]

HathiTrust Research Center: Pushing the Frontiers of Large Scale Text Analytics

J. Stephen Downie (University of Illinois Urbana-Champaign)

This presentation will introduce the newly-formed HathiTrust Research Center (HTRC) and its projects. The HTRC is affiliated with the HathiTrust, an online repository dedicated to the provision of access to a comprehensive body of published works for scholarship and education (http://www.hathitrust.org). Over 60 universities belong to the HathiTrust community and over 10 million volumes have been ingested into its digital archive from sources including Google Books, member university libraries, the Internet Archive, and numerous private collections. The HathiTrust Research Center (http://www.hathitrust.org/htrc) is dedicated to facilitating scholarship using this enormous corpus through enabling access to the corpus, developing research tools, fostering research projects and communities, and providing additional resources such as enhanced metadata and indices that will assist scholars to more easily exploit the HathiTrust corpus. The poster will outline the mission and goal of the HTRC, progress toward this goal to date, current and planned projects, and ways in which scholars can work with and through the HTRC.

Specific issues and related projects that will be discussed include:

(1) Development of approaches and tools for “non-consumptive research,” research on corpuses of documents that does not involve reading the documents. Of the 10.2 million volumes in HathiTrust, 62% are subject to copyright laws of the United States and other countries and the remaining 38% that are in the public domain are subject to Google terms of access. HTRC has entered into legal agreements with HT and Google that can provide computational access to the volumes as a service to researchers. In addition, HTRC researchers are currently developing computational tools for non-consumptive research that provide ways in which text mining and other tools can be applied to copyrighted materials of the HathiTrust corpus in ways that does not violate fair use terms.

(2) Development of a tool suite for analysis of HathiTrust data through collaborations with the SEASR group and Project Bamboo.

(3) Access to high performance computing resources for analysis of HathiTrust data.

(4) Specific HTRC research project examplars.

(5) Governance and community building.

<Keywords: Text, Text Mining, Digital Humanities, Non-consumptive Research, Google Books>