JADH2015

The Japanese Association for Digital Humanities 2015

Encoding Cultural Resources

文化資源をエンコードする

dates: Sep. 1-3, 2015

日程：2015年9月1日（火），2日（水），3日（木）

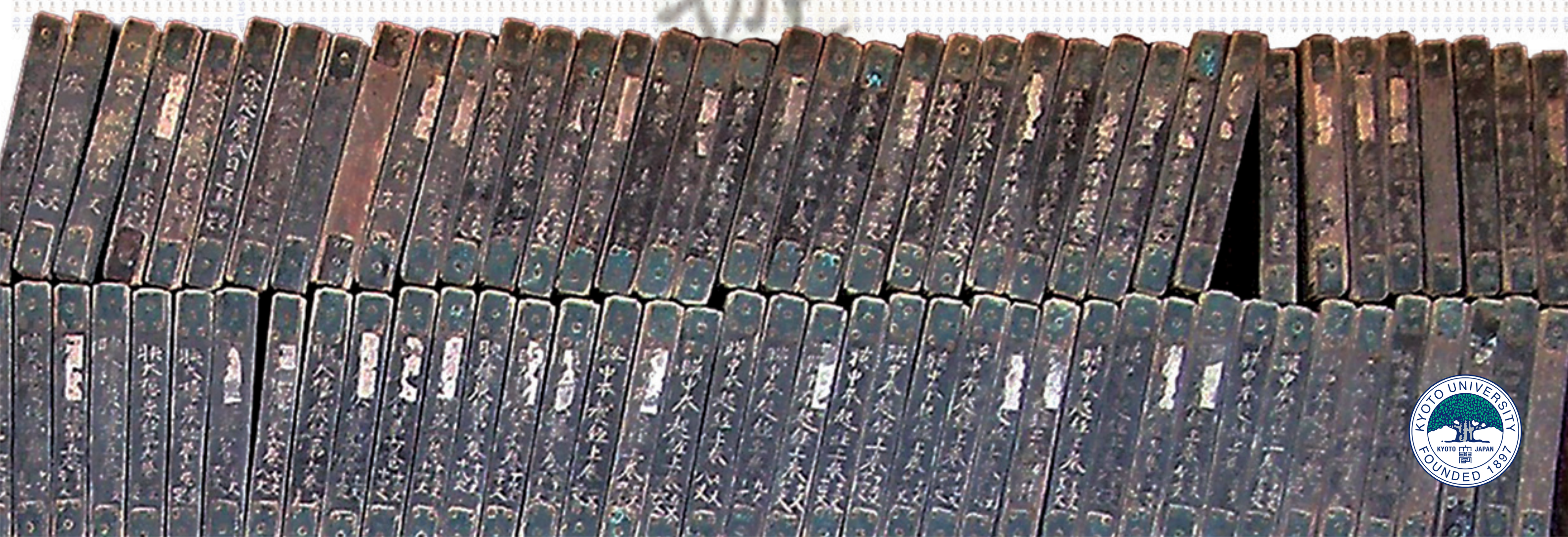
Kyoto University Institute for Research in Humanities

場所：京都大学人文科学研究所共通1講義室

Hosted by: JADH2015 Organizing Committee
Supported by: The Kyoto University Foundation

主催・問い合わせ先：JADH2015事務局
後援：京都大学教育研究振興財団

http://conf2015.jadh.org/



KYOTO UNIVERSITY
KYOTO JAPAN
FOUNDED 1897

JADH 2015

5th Conference of the Japanese Association for Digital Humanities

Encoding Cultural Resources

http://conf2015.jadh.org/

Conference Booklet

Kyoto University, Sep 1-3, 2015

Hosted by:

JADH2015 Organizing Committee [Members] under the auspices of the Japanese Association for Digital Humanities

Co-hosted by:

Institute for Research in Humanities (IRH), Kyoto University (KU)

Center for Informatics in East Asian Studies, IRH, KU

Center for Integrated Area Studies, KU

Graduate School of Letters, Kyoto University

Digital Humanities Initiative, Center for Evolving Humanities, Graduate School of Humanities and Sociology, The University of Tokyo

International Institute for Digital Humanities

Supported by

Kyoto University Research Administration Offices

The Kyoto University Foundation

Co-sponsored by:

IPSJ SIG Computers and the Humanities

Japan Art Documentation Society (JADS)

Japan Association for East Asian Text Processing (JAET)

Japan Association for English Corpus Studies

The Mathematical Linguistic Society of Japan

Program Committee:

Hiroyuki Akama (Tokyo Institute of Technology, Japan)
Paul Arthur (Australian National University, Australia)
James Cummings (University of Oxford, UK)
Neil Fraistat (University of Maryland, USA)
Makoto Goto (National Institute for Humanities, Japan)
Shoichiro Hara (Kyoto University, Japan)
Jieh Hsiang (National Taiwan University, Taiwan)
Asanobu Kitamoto (National Institute of Informatics, Japan)
Maki Miyake (Osaka University, Japan), **Chair**
A. Charles Muller (University of Tokyo, Japan)
Hajime Murai (Tokyo Institute of Technology, Japan)
Kiyonori Nagasaki (International Institute for Digital Humanities, Japan)
John Nerbonne (University of Groningen, Netherlands)
Espen S. Ore (University of Oslo, Norway)
Geoffrey Rockwell (University of Alberta, Canada)
Susan Schreibman (National University of Ireland Maynooth, Ireland)
Tim Sherratt (National Library of Australia, Australia)
Masahiro Shimoda (University of Tokyo, Japan)
Raymond Siemens (University of Victoria, Canada)
Keiko Suzuki (Ritsumeikan University, Japan)
Takafumi Suzuki (Toyo University, Japan)
Tomoji Tabata (Osaka University, Japan)
Toru Tomabechi (International Institute for Digital Humanities, Japan)
Christian Wittern (Kyoto University, Japan)
Taizo Yamada (University of Tokyo, Japan)

JADH 2015 Organizing Committee

Amano Eriko (Kyoto University)
Inaishi Natsuko (Kyoto University)
Kameda Akihiro (Kyoto University)
Kamiya Toshiro (Kyoto University)
Nagasaki Kiyonori (International Institute for Digital Humanities)
Tomabechi Toru (International Institute for Digital Humanities)
Wittern, Christian (Kyoto University) **Chair**
Yasuoka Koichi (Kyoto University)

Programme and Contents

Tuesday, September 1

Wednesday, September 2

09:30 Session 1: Structuring Data

11:20 Session 2: Textual Analysis (1)

14:00 Session 3: Culture and Digital Media

15:15 Break

15:35 Plenary Lecture 1:

16:35 Break

16:45 Poster/Demo Session

18:00 Banquet

Thursday, September 3, 2015

09:30 Session 4: Data Analysis

10:45 Break

11:05 Session 5: Textual Analysis (2)

12:20 Lunch Break

12:30-13:50 JADH Annual General Meeting

14:00 Plenary Lecture 2:

15:00 Break

15:20 Session 6: Research Infrastructures

17:05 – 17:30 Closing

"Gentle breath of yours my sails/ Must fill, or else my project fails":[1]

Tales of the Bodleian First Folio

In 1623, the First Folio of Shakespeare's plays was collaboratively published by a consortium of printers and fellow actors of the playwright, seven years after his death. A copy of the book arrived at Oxford's Bodleian Library that same year. It was deacquisitioned some time after the Third Folio's publication in 1663/4. Brought back to the Library by chance in 1905, a public funding campaign assured its permanent return to the Bodleian.

Due to its fragility, it has been little studied since. The goal of the book's second public campaign, *Sprint for Shakespeare*, was to create a "digital avatar"[2] which revealed the materiality of the book, made it accessible to readers primarily interested in the contents, and enabled reuse. A secondary aim of the campaign was to test the viability of crowd-funding digitization and online publication of rare books in the Bodleian Libraries' Special Collections, with a view to engaging the public with texts beyond the literary canon.

This talk tells some the book's curious stories, from its two public campaigns (1906 and 2012), through its conservation, digitization, and current publication in searchable text form, to the apple pip languishing in the pages of *Henry VI Part 2*. The remarkable story of perhaps the most-read copy of one of the world's most iconic books is a window on literary, social, and physical book history, and demonstrates some affordances of the digital for fans of Shakespeare, both scholarly and general.

This project was a collaboration across the Bodleian, the University of Oxford's e-Research Centre and IT Services, as well as faculty members from several universities. The resource was influenced by other digital facsimiles of early printed Shakespeare, including some of the team's previous work on the *Shakespeare Quartos Archive*. Work on the Bodleian First Folio continues.

Biography

Pip Willcox is the Curator of Digital Special Collections at the Bodleian Libraries, University of Oxford and responsible for its Centre for Digital Scholarship. With a background in scholarly editing and book history, she is an advocate for engaging new audiences for multidisciplinary scholarship and library collections through digital media. She conceived and ran the Sprint for Shakespeare public campaign and the Bodleian First Folio project. Current projects include Early English Print in the HathiTrust (ElEPHãT)—a linked semantic prototyping project, and SOCIAM: the theory and practice of social machines. Pip serves on the Text Encoding Initiative Board of Directors, and is Co-director of the annual Digital Humanities at Oxford Summer School, convening its introductory workshop strand.

[1] *The Tempest*, from *Mr William Shakespeare's Comedies, Histories, and Tragedies* (1623), line 2332 ('Epilogue, spoken by Prospero').

[2] S.M. Tarte. 'Interpreting Textual Artefacts: Cognitive Insights into Expert Practices.' In: *Proceedings of the Digital Humanities Congress 2012*, Sheffield, 2014. HRI Online Publications [http://www.hrionline.ac.uk/openbook/chapter/dhc2012-tarte].

Plenary Lecture 1 Wendell Piez: The Craft of XML

Profile

An American citizen whose father served with the US Department of State, Wendell Piez was born in Germany and grew up in Afghanistan (1963-1966), the Philippines (1966-1970), Virginia (1970-1975) and Japan (1975-1980), where he graduated high school from the American School in Japan (ASIJ), in Chofu, Tokyo. As a student at Yale College (1980-1984), he studied Classics, majoring in Ancient Greek. He received a PhD in English Literature from Rutgers University (1991), focusing on the history of literary form and writing his dissertation on the works of Walter Pater, an English scholar and critic (1839-1894) best known for his ideas of "aesthetic criticism" and his influence on literary Modernism.

At Rutgers University Special Collections and Archives and subsequently at the Center for Electronic Texts in the Humanities (CETH) based at Rutgers and Princeton Universities, Dr Piez began applying his skills in computer programming to problems in humanities scholarship. While maintaining professional contacts in this field, he left the academy in 1998 for work in the private sector. At Mulberry Technologies, Inc., a consultancy specializing in the design and application markup technologies (SGML and XML) especially in the scholarly and technical (not-for-profit) publishing industry, Dr Piez became known as a practitioner and instructor of XML and related technologies, including XSLT. Since 2012 he has been an independent consultant in private practice.

Dr Piez serves on the faculty at the University of Illinois (UIUC) Graduate School of Information and Library Science (GSLIS), and as General Editor of Digital Humanities Quarterly (DHQ - http://digitalhumanities.org/dhq/). He is a member of the Association for Computers in the Humanities (ACH) and of the Text Encoding Initiative (TEI). His papers on the theory and practice of markup technologies have been presented at industry and academic conferences including the Digital Humanities conferences, TEI Members Meetings, and Balisage: the Markup Conference. His professional web site is at http://www.wendellpiez.com.

Abstract

What would it mean to treat XML as a craft? In many ways, information technology is the opposite of a craft, since it is based in principles of automation and devoted to the idea of minimizing or eliminating labor, without which craft is impossible. Yet when we look carefully at this picture, things become more complex: even a brief look at the history of technology can show that as it develops, craft does not disappear even with automation. Blind machines, carrying out our orders, may be necessary for automated mass production; but it requires craft to build and use them.

XML in particular has a layered architecture: on top of "raw text", XML describes a syntax; a way of defining a vocabulary for a markup or data description language using that syntax; and ways of defining processing (application logic) for such a language. This organization reflects our needs both for standardization -- in order that we can share methods and tools for working with textual information -- and for customization and adaptability to our own peculiar problems. Because of this combination of strength and flexibility, XML is especially suitable for any work with digital text that requires and rewards the care and sensitivity of craftmanship -- when information is rare, expensive and valuable -- such as projects in the digital humanities. A small demonstration, HaiKuML (a platform for the study of Japanese poetry by students of the language), presents an opportunity to see how this might be.

HaiKuML is on the Internet at https://github.com/wendellpiez/HaiKuML, where it may be downloaded for trial and testing.

Plenary Lecture 2   Øyvind Eide:
The Text is Not the Map – Using Conceptual Modelling to Understand the Non-Mappable Aspects of Narrative

Profile

Øyvind Eide holds a PhD in Digital Humanities from King's College London (2012) which was funded by a grant from The Research Council of Norway. He was been an employee in various positions at The University of Oslo from 1995 to 2013, most recently as a Senior Analyst at The Unit for Digital Documentation. From October 2013 he is a Lecturer and research associate with the Chair of Digital Humanities at The University of Passau.

Eide is an elected member of the board of The European Association for Digital Humanities (EADH) from 2008 to 2014 and a co-opted member from 2014 to 2016. He is the Chair of the Awards Committee The Alliance of Digital Humanities Organizations (ADHO) from 2011 to 2017. He is an elected member of the executive of the ADHO SIG Global Outlook::Digital Humanities (GO:DH) from 2014 to 2016.

His research interests are focused on conceptual modelling of cultural heritage information as a tool for critical engagement with the media differences, especially the relationships between texts and maps as media of communication. He is also engaged in theoretical studies of modelling in the humanities as well as beyond.

Abstract

The relationship between texts and maps in light of interart and intermedia studies will be the topic of this paper. Are we forced to tell different stories using the two media? The process of creating maps based on texts will be connected to differences between the two media at a formal level. Transformative digital intermedia studies will be introduced together with its core method: critical stepwise formalisation. It will be shown how critical stepwise formalisation as a computer assisted conceptual modelling method is well suited to study media differences at a micro level.

While it will be documented how several types of textual expressions are not translatable to maps, it will not be claimed that the two should be separated. On the contrary: because of their different sign systems and because they can present different if overlapping views of the world, combining them is necessary to move towards richer geographical stories.

Landscape descriptions taken from two texts from the eighteenth century will be presented as case

studies in the paper. One is a non-fiction text written as background material for border negotiations in Scandinavia in the early 1740s (Schnitler 1962), whereas the other is Defoe's novel Robinson Crusoe from 1719 (Defoe 2008). A conceptual model based on the former text was used in critical stepwise formalisation to test how the spatial understanding expressed in the text could be expressed as maps. Thus, comparisons with maps, specific ones as well as a generalised idea of the Map, put the model in perspective.

Through the application of the results to a reading of Robinson Crusoe, new understanding of Schnitler as well as of Defoe was developed. Maps can be seen as a type of image, so the reflection upon differences between verbal texts and maps brings us into the area of comparison between text and image as well. The Ut pictura poesis tradition is therefore brought into the discussion, using some points taken from another eighteenth century text, namely, Lessing's Laokoon from 1766 (Lessing et al. 1893). In order to put the discussion into the perspective of landscape description in different media, some modern theory in the area of map semiotics (MacEachren 2004) and media modality (Elleström 2010) will also be used.

The paper will continue with a discussion of a contemporary text where the landscape being described is surreal and dreamlike, namely, Ishugoru's 1995 novel The Unconsoled. How do the thinking developed above work for such a text? Will it break down, or can it still be used to say meaningful things about the textual landscape? And what would it mean to develop maps based on such a text? Through this it will be demonstrated not only how intermedia studies can be enriched by the use of critical stepwise formalisation, but also how the method is useful to understand better how texts create virtual landscaped in the mind of the reader.

References

Daniel Defoe. The life and strange surprizing adventures of Robinson Crusoe (1719). Volume 1 of The novels of Daniel Defoe. London: Pickering & Chatto, 2008.

Eide, Øyvind. Media Boundaries and Conceptual Modelling : Between Texts and Maps. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan, Forthcoming 2015.

Lars Elleström. “The Modalities of Media: A Model for Understanding Intermedial Relations.” Media borders, multimodality and intermediality. Ed. Lars Elleström. Basingstoke: Palgrave McMillan, 2010. – P. 11–48.

Kazuo Ishiguro. The Unconsoled. London: Faber and Faber, 1995.

Gotthold Ephraim Lessing, Karl Lachmann, and Franz Muncker. Gotthold Ephraim Lessings sämtliche Schriften. Neunter Band. Photographic reprint. Berlin : De Gruyter, 1968. 3., aufs neue durchges. und verm. aufl. edition. Stuttgart: Göschen, 1893.

Alan M. MacEachren. How maps work : representation, visualization, and design. First published in 1995. New York: Guilford Press, 2004.

Peter Schnitler. Major Peter Schnitlers grenseeksaminasjonsprotokoller 1742–1745. Volume 1. -Oslo: Norsk historisk kjeldeskrift-institutt, 1962.

MemoryHunt: A Mobile App with an Active Viewfinder for Crowdsourced Annotation through the Re-experience of the Photographer

Asanobu Kitamoto

A photograph is usually considered as the record of the world, but this definition misses the important element of a photograph, namely the photographer. When a photograph is treated as a historical resource, it is important to interpret where, how, and why the photograph is taken. These kinds of information could be inferred from simulating the photographer in terms of location, direction, and posture. Re-experiencing the photographer is the design goal of our mobile app, “MemoryHunt.”

The basic idea of the app is simple. An old photograph is shown on the viewfinder of the camera with adjustable transparency to allow users direct comparison between the old photograph and the current landscape. Users then move around to find the best match between them, and pressing a shutter button records both the current landscape and metadata of the camera such as location by GPS and direction by motion sensors. This amounts to annotating the old photograph with coordinate information and the temporal change of landscape. Because the app can be distributed to anyone from the app market, the app has potential to be deployed into crowdsourced annotation of photographs based on field work by massive people.

The mechanism of the app is simple, but the concept of the app is tricky to understand, especially in terms of differentiating this app from augmented reality (AR). Some of the app users got confused by superficial similarity with AR, because both approaches focus on the matching of photographs with the real world. A fundamental difference is in the “screen overlay” approach of the app in contrast to the “ground overlay” approach of augmented reality. The photograph is fixed on the viewfinder in the former, while the photograph is fixed on the real world in the latter. Because of this difference, augmented reality is a technology for exhibition, but MemoryHunt is a technology for participation. In the former, users can just confirm the appearance of the photograph that was already prepared by someone else, but in the latter, users explore the real world to discover the best match. Exploration of the real world is fun, because users can check in real-time if their body motion improved the accuracy of matching or not. This feedback mechanism, with the appropriate level of challenge, is known to be the essential part of so-called “flow” and the sense of fun naturally leads to gamification mechanism by transforming photo-taking into an enjoyable experience by offering achievable challenges.

We propose that the fundamental concept of MemoryHunt is “active viewfinder” in the sense that the app changes a viewfinder into a place for suggesting the next photograph to take. The opacity of viewfinder is the key to take a simulated photograph, on the contrary to a transparent viewfinder that let photographers to take anything. Extending this concept, it is clear that active viewfinder can show not only old photographs but also any other visual materials such as paintings, animes, movies, and album covers. People can enjoy taking the same scene with an existing masterpiece from the same place and the same direction, and the collection of these photographs, in turn, can be considered as crowd-based pseudo-fixed-point observation. We imagine that this will evolve into a new photographic culture in which photographs are linked to the database of photographs.

We released MemoryHunt at Google Play, and held five workshops to observe the actual usage of the app. Two of them at Tokyo (Japan) and Kyoto (Japan) used historical photographs, but three of them at Aceh (Indonesia) and Kobe (Japan) used photographs before or just after the disaster. This is because MemoryHunt can uncover not only the long term change of landscape in history, but also the short term change of landscape from destruction by the disaster to recovery from the disaster. At Aceh, Indonesian local people quickly understood the app and enjoyed exploring the change of landscape in ten years after Tsunami in 2004. At Kobe, children also enjoyed exploring the town to compare the photograph taken in 1995, which is before their birth.

Lastly, crossing the border of time by MemoryHunt could become an emotional experience. To take the same photograph, a user needs to stand at the same place, and simulate the same posture with the photographer. Moreover, the user might need to simulate the emotion of the photographer about why the photograph was taken in this way. As we deepen the interpretation of the photograph, re-experiencing the photographer becomes an emotional experience to match yourself with the photographer across time.

Acknowledgment: The project was supported by Center for Integrated Area Studies (CIAS), Kyoto University. The app was developed by a software engineer, Mr. Tomohiro Ikezaki. An experiment at Aceh, Indonesia, was performed by Dr. Yoshimi Nishi and Dr. Hiroyuki Yamamoto at CIAS. Other credits are summarized in the project website, http://dsr.nii.ac.jp/memory-hunting/

Making Links: Connecting humanities resources for scholarly investigation

Kevin Page

Expressing relationships between items in corpora using Linked Data provides a powerful means for reusing knowledge from one domain or dataset when investigating another. The process of forging those links can, however, be complex and time consuming – a careful balance must be struck between assistive technologies that can automate elements of the method so enabling the researcher to work at scale, while oversight and clarity must be maintained for those links that encapsulate scholarly judgements. We report our experience creating such links with investigative scenarios in music and early English texts: of creating and publishing the RDF foundations, of the ontological structures that encode the relationships, of our Semantic Alignment and Linking Tool that assists the investigator in the process of finding matches, and of the scholarship that is built upon the tools and data.

Our first case study, the ElEPHT project - Early English Print in HathiTrust, a Linked Semantic Worksets Prototype - demonstrates the use of Linked Data for combining, through worksets, information from independent collections into a coherent view which can be studied and analyzed to facilitate and improve academic investigation of the constituents. The project focuses on the symbiosis between two datasets: the first is Early English Books Online - Text Creation Partnership (EEBO-TCP), a mature corpus of digitized content consisting of English text from the first book printed through to 1700, with highly accurate, fully-searchable, XML-encoded texts; the second is a custom dataset from the HathiTrust Digital Library of all materials in English published between 1470 and 1700. The project addresses a number of challenges which we present here:

1. generating RDF metadata from EEBO-TCP to complement the HathiTrust RDF;
2. identifying suitable ontologies for encoding the EEBO-TCP RDF that can be usefully linked to the HathiTrust data, and to other external entities;
3. identifying and aligning co-referenced entities within both datasets, and storing these as RDF;
4. providing infrastructure to host the RDF datasets and SPARQL query interfaces;
5. creating SPARQL queries of sufficient expressivity to parameterise worksets for scholarly investigation;
6. demonstrating the construction and utility of such parameterised worksets through prototype user interfaces, showing how a researcher can create and view a workset and the content within it.

At this time the ElEPHT datastore contains 1,137,502 triples, consisting of: 251,725 entities, 66 distinct classes, 214 distinct predicates. 287,581 distinct subject nodes and 294,677 distinct objects. Our second case study considers data sources in music, firstly in terms of digitised images and metadata from the British Library Early Music Online (EMO) linked with the Electronic Corpus Of Lute Music (ECOLM), and secondly in the linking and scholarly use of data originating from commercial and more generic sources including the BBC, Musicbrainz, DBpedia and VIAF. In this context we present the Semantic Alignment and Linking Tool (SALT) that addresses the problem of creating inter-dataset links by applying semantic technologies and Linked Data approaches in order to produce candidate alignment suggestions that may be confirmed or disputed by the humanist. These decisions are integrated back into the knowledge base and are available for further iterative comparison by the user; the complete RDF graph is published and can be queried through the same SPARQL endpoint that also underlies the SALT user interface. Provenance of the musicologist's judgement is captured and added to the descriptive graph, supporting further discourse and counter-proposals. We report on the design and implementation of this tool and its evaluation within this musicological context.

Finally, we reflect on common patterns and lessons that emerge from these scenarios, and what they might tell us about the nature of humanities scholarship with broadly sources, dynamic, and interlinked data. In doing so we accept and reinforce the observation that the act of study is iterative and ongoing; that the linking tools can provide both a means for capturing the provenance of judgements over complex information structures, and of incorporating these judgements in new and dynamic data structures that can, in turn, provide the foundation for further insight.

Digital Editions and Citing Them – Making scholar friendly digital editions

Espen S. Ore

A text edition – scholarly or documentary – establishes a text that ideally will be quoted and cited by other scholars. A printed edition is a dead edition in the meaning that it is fixed. A digital edition can be modified, and this is both a strength and a weakness. Without going into the discussion of what a text is, it is important for me that if I cite a draft version of Henrik Ibsen's play A Doll's House where I mention the clear symbolism in the male protagonist's name "Stenborg" ("Stone castle") which was changed in the final version to "Helmer" I would be very unhappy if the digital edition1 of this draft manuscript was emended to use the final version of the name.

The non-fixed state of digital web-accessible editions was realized many years ago. Different theoretical and practical solutions to the problems that come from this fluid state of the texts have been proposed and to a certain degree implemented. But these solutions are not universal, and they look at the problem from different points of view: some solutions seem to come from a librarian or bibliographic such as the Chicago Manual of Style. Here we fins that the style takes care to note when a digital edition has been accessed and if possible a modification date for the resource cited.2

The systems for citing are tools on the reader's side of digital publication. There are also possible tools on the publishing side that may be implemented to a certain degree. Among the technical solutions is the use of check sums and other ways of computationally checking the content. In addition it is possible to mark a version by giving it some sort of unique identifier. These identifiers such as an ISBN or a DOI are not in themselves linked to a certain URI - in principle the content identified by for instance a DOI can be moved between servers as long as a reference system keeps track of the recent valid URI for the document. But when it comes to the real world and academic publications of digital editions things might not be that easy. If we look at for instance the digital edition of Henrik Ibsen's Writings, the current URL for the edited scholarly critical edition of A Doll's House is:
http://www.ibsen.uio.no/DRVIT_Du%7CDuht.xhtml

This links to the text in version 1.1 of the digital edition. There will of course be links to version 1.0, but existing links directly to the text will from now on fetch texts as they are in the 1.1 version, not in the 1.0 version (if there are differences). And when a version 1.2 or 2.0 is launched, the link quoted above will refer to thsi new version. This may be a problem. The solution that has been selected for the digital Ibsen is to include a link to iformation about the different versions at the bottom of the page labelled "Endringer fra tidligere versjoner". Clickong on this link leads to a page with nformation about the changes between the versions and maybe just as important, a further link to te earlier version(s). This still means that the link to the edition of Ibsen's play quoted above will always refer to the newest version, and so the link used as a quote in for instance a scholarly paper may be wrong unless the quote also cotains information about the version numer. This is not without problems for the user. But a little cursory reading through web pages from various digital editions seem to show that this is a universal problem, not one particular to the Ibsen edition. It may also be a difference between digital editions published for the general public and following (more or less) the structure of a printed version and editions that are built upon the model of a paper archive.

In this paper I will compare the solutions (or lack of solutions) found in various large scale digital editions produced in the Nordic countries and ideally nominate a list of the solutions that seem to handle this problem best. The editions I will look at are works by N.F.S. Grundtvig, Ludvig Holberg, Henrik Ibsen, Søren Kierkegaard, Selma Lagerlöf, Edvard Munch, August Strindberg and Zakarias Topelius while I will compare those to the Medieval Nordic Text Archive (MENOTA) and the Wiitgenstein Archives at the University of Bergen (WAB).

Towards Automatic Extraction of Punch-lines: Plot Analysis of Shinichi Hoshi’s Flash fictions

Hajime Murai

Introduction

Through information processing technologies developed in recent years, many quantitative analyses of literature, including bibliometrics, have been carried out in various ways. Though it is difficult to capture all story structures and meanings using machines at present, it is possible to incorporate quantitative indicators of narrative analysis in order to enhance the objectivity of the story analysis. Focusing on this outcome, utilizing an eclectic approach of quantitative and traditional humanities methods, such as structural analysis [1] and conventional plot analysis [2], the characteristics of narrative structure and changes in the narrative pattern can be extracted [3]. Moreover, the structures of punch-lines can be detectable by focusing on the behaviors of the characters in the story [4].

The purpose of this research is to expand a plot analysis method suitable for describing parallel story lines and punch lines. If it were possible to describe punch lines in complex stories, a database suitable for capturing general narrative structure could be realized. In addition, a method for automatic punch-line extraction is the goal of this research. In this research, the works of flash fiction by Shinichi Hoshi were analyzed as a case study and prototype JAVA program were developed.

Plot and Behavior of Characters

When the characters’ behaviors are classified and registered in a database, it will be useful to ascribe attributes to them in order to compare their functions in similar stories. For example, if the relationship of the protagonist and antagonist is hostile, the representation of hostility can take various forms, such as the destruction of property, reputational damage, or physical harm. However, it is desirable that those various representations be categorized in a unified manner.

In order to identify the changes in plot structure, it would be useful to include the relationship of agent (subject person of the behavior), patient (object person of the behavior), and the target of certain behaviors in the database. This method of analysis is particularly applicable to instances of situation reversal in the narratives that characterize the punch lines of Hoshi’s short-short stories.

Therefore, in drafting this study, the behaviors of the characters are categorized by three attributes: focus, polarity, and person. The focus is the type of behavior. The polarity is the negativity and positivity of the effect of the behavior. The person is the relationship of agent character and patient character in the story. Moreover, the motivation of the behavior and its results are described in accordance with these three attributes in order to record the plot data before and after the behavior.

From the works of flash fiction by Shinichi Hoshi, three major genres were analyzed (Stories about universe, devils and drugs. Those encompass a total of approximately 200 stories) from the view point of the characters’ behaviors.

Extraction of Punch-Line

Based on the results of plot analysis, patterns of the punch-lines in Shinichi Hoshi's works were manually extracted. Most of those include some reversal pattern as punch-lines. Such as reversal of agent-patient relationship, trade-off reversal, evaluation reversal, common sense reversal, purpose reversal and so on [4]. Within those punch-lines related to reversal, reversals of role of characters and reversals of behaviors can be extracted comparatively easily from the result of plot analysis. Therefore, prototype JAVA program to extract punch lines based on reversals of role of characters and behaviors were developed. This prototype JAVA program analyzed data of plot structures and approximately 60% of punch-lines in Shinichi Hoshi's works were extracted as patterns of reversal. However, the results of prototype JAVA program also include reversals which are not punch-lines. Therefore it is necessarily to select valid punch-line patterns from all reversal patterns.

Rules for Selecting Valid Punch-Line

In order to evaluate validity of punch-line candidates, the general common features of the punch-lines which are composed of reversal pattern of plot elements were examined based on the result of former plot analyses. As a result, three features (convergence, similarity, and distant foreshadowing) were extracted.

Convergence: The plot element which relates multiple reversal patterns has high possibility of being a valid punch-line.

Similarity: Similarity between behaviors in the two plot elements which compose the reversal pattern increases the possibility of being a valid punch-line.

Distant Foreshadowing: If the plot elements which compose reversal pattern locate distantly or locate near the end of the story, the possibility of being a valid punch-line becomes high.
Implemented those features as an algorithm, it became possible to sort reversal patterns and to pick up valid punch-lines in most of Shinichi Hoshi's works.

Conclusions and Future Works

In this research, the works of flash fiction by Shinichi Hoshi were analyzed as a case study based on the plot description which focuses on the behaviors of the characters in the story. In order to enable automatic punch-line detection, prototype JAVA program were developed. Moreover, it is clear that convergence, similarity and distant foreshadowing are common tendencies for valid punch-lines.
In this research, target works of the plot analysis were limited and the result of the plot analysis was not double checked. Therefore it is desirable to extend this analysis to other works and to confirm the validity of methods in this research. Moreover, the plot analysis should be done by multiple researchers and the result should be checked objectively.

References

1. Roland Barthes, Elements of Semiology, Hill and Wang, New York, USA, 1968.
2. Vladimir Propp, Morphology of the Folk Tale, U of Texas P, 1968.
3. Hajime Murai, Naoko Matsumoto, Chie Sato, and Akifumi Tokosumi, Towards the numerical analysis of narrative structure: The characteristics of narrative structure within the short-short stories of Shinichi Hoshi, Journal of Japan Society of Information and Knowledge, Vol. 21, No. 1, pp. 6-17, 2011.

4. Hajime Murai, Plot Analysis for Describing Punch Lines in Short-Short Stories, OpenAccess Series in Informatics, Vol. 41, pp. 121-129, 2014.

Comparing the Addresses of Russian Presidents in Inaugural Ceremonies: A Text-mining Approach

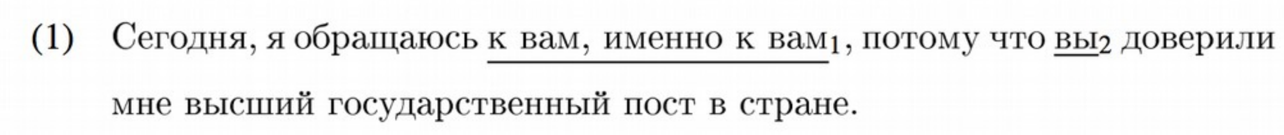
Sugiyama Mao

This study proposes that a close observation in personal pronouns in speeches given in an inaugural ceremony by Russian Presidents and ex-presidents can be of help to understand the political intentions of presidents.

Russian Presidential speeches have received great attention. For example, Ueno (2009) reveals that an annual state-of-the-nation address given by Medvedev in 2008 is drastically different from addresses by other presidents in that speeches byYeltsin and Putin make no explicit references to a political reform, while Medvedev's address does. Ishino (2014) observes an annual state-of-the-nation address given by Putin in 2014. Being persuasive, these studies put a focus on how the presidents speak about, and not what they say. This study suggests that a linguistic perspective brings a new insight to a speech study. Kotynya (2011) compares the speeches Putin's and other candidates' before election in 2000. Furthermore Filinskiy (2010) makes a point that Putin expresses an opinion by using vulgar language.

This analysis deals with speeches given in an inaugural ceremony. A new Russian president and an ex-president give the speeches at the ceremony. Addresses as a new president included in this study are Putin's given in 2000 and in 2012 and Medvedev'sin 2008, those as an ex-president dealt areYeltsin's in 2000, Putin's in 2008, and Medvedev's in 2012.

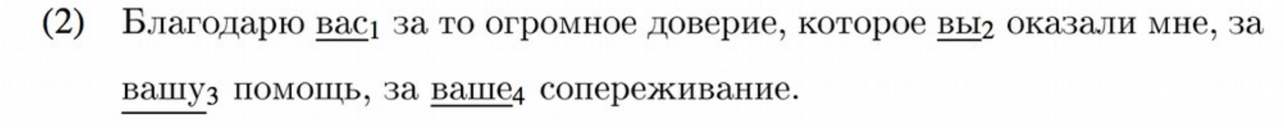
The purposes of the speeches are to express presidents' policy, hopes, thought, etc. to citizens. Hence, the addresses should contain expressions that show either hidden or intended attitudes of presidents toward citizens.We discuss that the use of personal pronouns, especially how each president uses MbI ('we') and BbI ('you'), A sense of distance with nations that the presidents have. For example, in the case of BbI ('you'), in 2000, Putin frequently uses vi to emphasize that it is the citizens who elected him as a president – not authorityofYeltsin – as demonstratedin (1).



(1) Сегодня, я обращаюсь к вам, именно к вам$_1$, потому что вы$_2$ доверили мне высший государственный пост в стране.

'Today, I speak to you, that is to you1 because you2 have entrusted me with the highest office in the country.' (Putin, 2000)

Contrary to Putin, Medvedev uses BbI ('you') to show his honor to citizens as in (2).

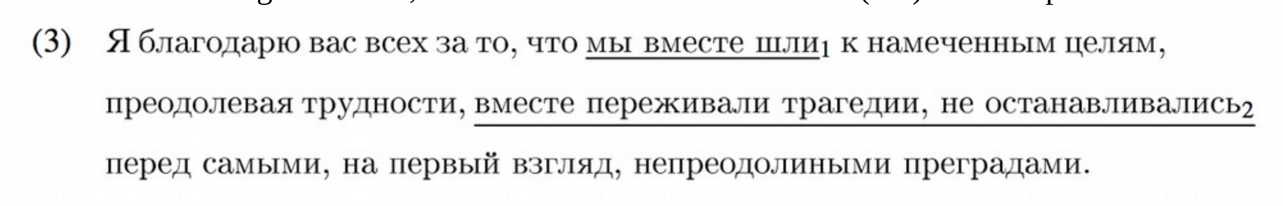


(2) Благодарю вас$_1$ за то огромное доверие, которое вы$_2$ оказали мне, за вашу$_3$ помощь, за ваше$_4$ сопереживание.

'Thank you1 for the great confidence you2 have given me, for your3 help, for your4 compassion.' (Medvedev, 2012)

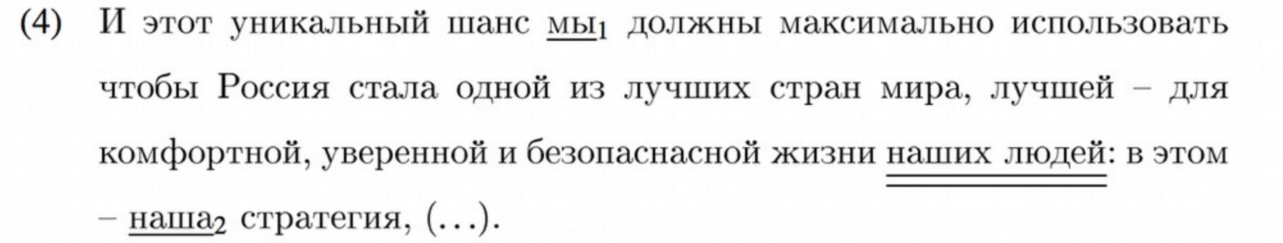
These two usages should be distinguished because Putin's usage shows a locus of responsibility, whereas Medvedev's one shows his respect to nations.

Second, the reference of first plural person pronouns shows a different attitude toward citizens. On the one hand, Putin uses MbI (‘we’) referring to citizens and Putin himself : on the other hand, Medvedev refers to government, and does not include nations. See (3-4) for examples.



(3) Я благодарю вас всех за то, что мы вместе шли$_1$ к намеченным целям, преодолевая трудности, вместе переживали трагедии, не останавливались$_2$ перед самыми, на первый взгляд, непреодолиными преградами.

‘I would like to appreciate all of you for heading toward goals together1, suffering and overcoming difficulties, and making progress together2 in spite of seemingly insurmountable obstacles.’ (Putin, 2008)

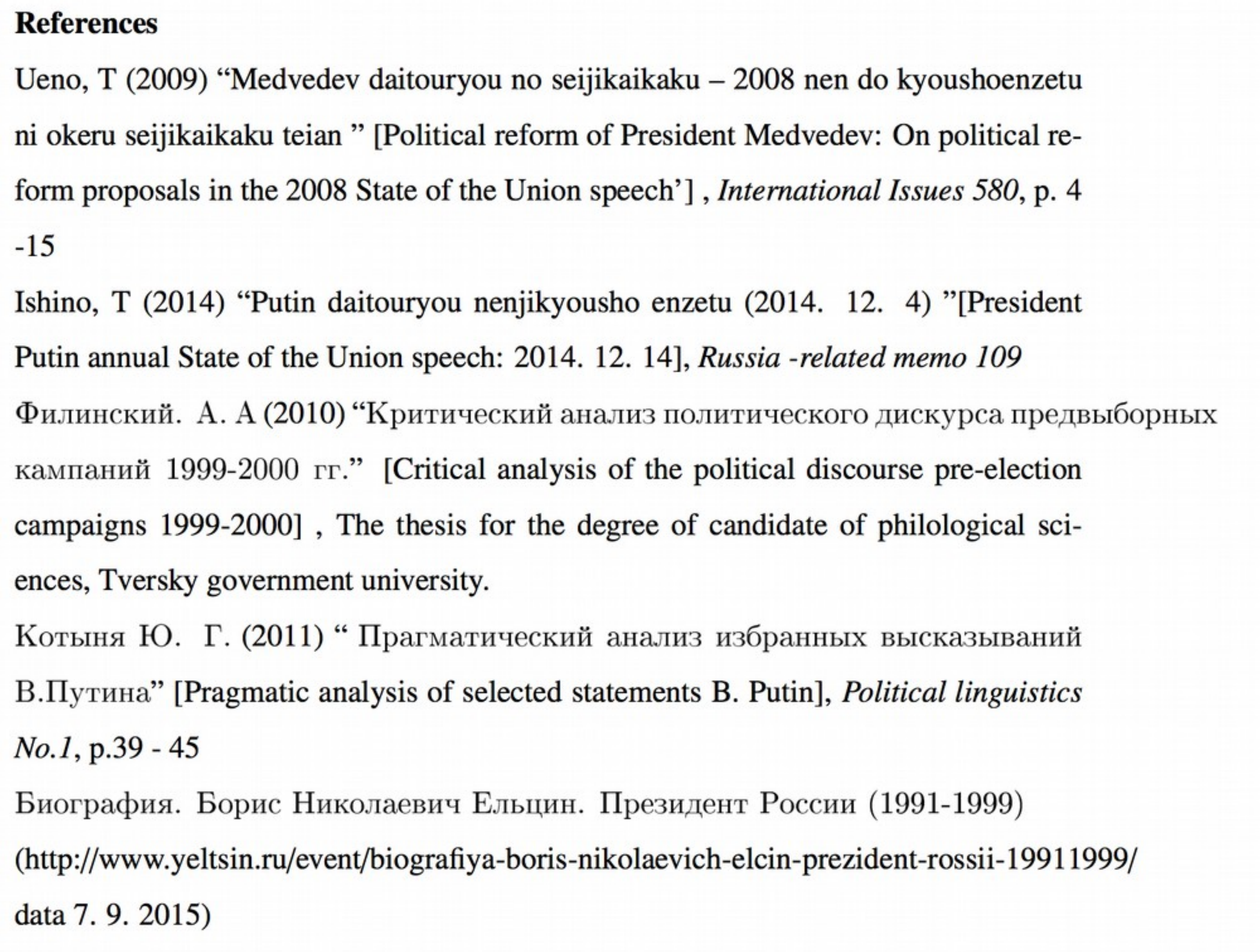


(4) И этот уникальный шанс мы$_1$ должны максимально использовать чтобы Россия стала одной из лучших стран мира, лучшей – для комфортной, уверенной и безопаснасной жизни наших людей: в этом – наша$_2$ стратегия, (...).

‘We1 have to make the most of this unique opportunity to make Russia be one of the best countries in the world, better country –more comfortable, more confident and more secure life of our people –which is our2 strategy, ’ (Medvedev, 2008)

In(3),as theword ‘together’ indicates,Putin tries to get people into his camp by seeing goals and challenges in the same way as citizen does. Contrary to Putin, Medvedev draws a clear line between the government and the public. This can be observed in ‘our people’, indicating that people belong to the government. This difference in referent in MbI (‘we’) suggests that whether a president considers himself as a member of nation or as a person with a different status reflects presidents’ attitude toward citizens and political measures.

This study investigated speeches at the Inaugural Ceremonies made by a new Russian President and ex-president. Although previous studies are insightful, linguistic perspectives have been neglected. The current study demonstrates that the use of pronouns reveals a mental attitude toward listeners that a speaker or a lecturer potentially have, or intend to show. This aspect may be overlooked in approaches taken by the previous studies cited in this paper.



References

Ueno, T (2009) “Medvedev daitouryou no seijikaikaku – 2008 nen do kyoushoenzetu ni okeru seijikaikaku teian ” [Political reform of President Medvedev: On political reform proposals in the 2008 State of the Union speech’] , *International Issues 580*, p. 4 -15

Ishino, T (2014) “Putin daitouryou nenjikyousho enzetu (2014. 12. 4) ”[President Putin annual State of the Union speech: 2014. 12. 14], *Russia -related memo 109*

Филинский. А. А (2010) “Критический анализ политического дискурса предвыборных кампаний 1999-2000 гг.” [Critical analysis of the political discourse pre-election campaigns 1999-2000] , The thesis for the degree of candidate of philological sciences, Tversky government university.

Котыня Ю. Г. (2011) “ Прагматический анализ избранных высказываний В.Путина” [Pragmatic analysis of selected statements B. Putin], *Political linguistics No.1*, p.39 - 45

Биография. Борис Николаевич Ельцин. Президент России (1991-1999) (http://www.yeltsin.ru/event/biografiya-boris-nikolaevich-elcin-prezident-rossii-19911999/ data 7. 9. 2015)

An Effect-Size Analysis of Christianity in the 19th-Century British Novels

Tatsuhiro Ohno

This paper applies “effect size”—a corpus-linguistic technique for highlighting keywords of a study corpus by comparison with its reference corpus--to the best objective analysis of contrast among the nine eminent 19th-century British novelists in their depictions of Christianity.

It is universally acknowledged that reading is “an activity of a subjective nature” (Bal 4). Critics’ instinct provides us with insightful readings of texts which have made notable contribution to the enrichment of our present life. One of the serious problems of such subjective readings, however, would be that it is almost impossible to pinpoint the true interpretation of a literary text (Sabol 47).

Many critics have argued that there is no absolute interpretation of a text which is hardly identifiable even by the author him/herself. Terry Eagleton, a typical anti-intentionalist critic, insists that we “can never . . . come to know in some absolutely objective way” what an author has actually in mind, and that any “such notion of absolute objectivity is an illusion” (60). Nowadays, in addition, few critics seem to endorse the assertion of intentionalist hermeneuticists who consider a literary text has “one and only one correct interpretation” (Juhl 13, 238) that “Any interpretation that exactly and completely captures the author’s intended communication would be a definitive interpretation” (Irwin 62).

My research takes a structuralist approach, or applies corpus linguistic techniques—statistics, concordance, effect size, and topic modelling—to the interpretation of literary texts. This so-called “corpus stylistic” method looks one of the most innovative means that could contribute to the objective discovery of a definitive interpretation of a text, since it discloses the hidden structures of a text which are invisible in traditional readings of a text, and therefore could be closely interlaced with its authorial meaning. The purport of this computer-assisted method is succinctly summarized in David I. Holmes’s statement: “The statistical analysis of a literary text can be justified by the need to apply an objective methodology to works that for a long time have received only impressionistic and subjective treatment” (18).

Effect size, which is the prime approach taken in this paper for this purpose, is the “% difference of the frequency of a word in the study corpus when compared to that in the reference corpus” (Gabrielatos and Marchi 9), or a means for detecting frequency difference between the study corpus and its reference corpus. It can be calculated by the following formula (Gabrielatos and Marchi 12):

The study corpus of nine novelists--Jane Austen, Elizabeth Gaskell, Charles Dickens, Charlotte Brontë, Emily Brontë, George Eliot, Anne Brontë, Thomas Hardy, and George Gissing—is compared with the reference corpus of the 291 novels by the 28 eminent 19th-century British authors. The software used is AntConc; the e-texts are principally taken from Project Gutenberg. Data calculation and graphic production are conducted on Microsoft Excel.

The top-rank effect-size keywords used by the nine novelists are classified into 15 categories, including GOD_morality, GOD_nature, HUMAN BEINGS_action, HUMAN BEINGS_people, LANGUAGE_dialect, PLACE_buildings, and TIME, according to the connotation or context of each word. As Otto Jespersen confesses, it may be sometimes difficult “to find a satisfactory classification of all the logical relations” (137), but, even so, classification reveals the distinctive choice of words by each author construct.1 The result is demonstrated in Fig. 1.

The focus is placed on the novelists’ descriptions of Christianity chiefly for the purpose of scrutinizing the 19th-century British people’s response to God’s Plan of Salvation, or eternal life—one of the key principles of the Christian creed—, and thus inspecting the truthfulness of the teaching. Hence, the effect-size values of the God-related words only are extracted from the result above, and visualized in Fig. 2.

Especially conspicuous in this outcome would be the features of the following ten.
(a) First of all, the effect-size sores of the nine novelists’ use of God-related words are significantly lower than those of their use of other keywords.
(b) No detectable use of God-related words is found in the novels of Dickens and E. Brontë, although their interest in Christianity is one of the popular subjects of critics’ discussion.
(c) Hardy’s use of nature-related words (e.g. “heath,” “oak,” and “boughs”) is by far the most frequent among the nine authors’, and C. Brontë’s (e.g. “garden,” “flowers,” and “moon”) the second most.
(d) The effect size of G. Eliot’s moral-related words exhibits the decisively high score of more than 4,700 points, among which the top five out of the 13 words are concerned with preachers of the Christian doctrine, such as “frate,” “preaching,” “rector,” “preacher,” and “vicar.”
(e) The second-rank utilizer of the words of moral connotation is Gaskell, whose effect-size keywords in the category are “thankful,” “tender,” “sin,” “comfort,” and “bless.”
(f) The third-rank user is Austen, and her effect-size keywords are “comfort,” “affection,” and “kindness.”
(g) Effect-size inquiry into Gissing’s novels discloses only one word of divinity-associated keyword, “sincerity.”
(h) C. Brontë displays three moral-related words of high effect-size value: “pure,” “affection,” and “spirit.”
(i) The only effect-size keyword of the moral connotation among Hardy’s 22 God-connected keywords is “vicar.”
(j) The only divinity-associated keyword of A. Brontë that appears in the effect-size analysis of her fiction is “god.” A contextual analysis of the word unearths that her protagonists are given the role of spokesperson of the author’s Christian faith.
The effect-size analysis of Christianity in the 19th-century British novels brings to light some latent differences and similarities among the nine renowned authors in their descriptions of divine elements which are hardly identifiable by conventional reading of a text. The individual author’s degree of commitment to God’s Plan of Salvation will be explored by quoting evidence from the texts—an instance of the combination of distant and close readings. The outcome will be explained in my presentation.

Works Cited

Bal, Mieke. Narratology: Introduction to the Theory of Narrative. 2nd ed. Toronto: U of Toronto P, 1997. Print.
Eagleton, Terry. Literary Theory: An Introduction. 2nd ed. Oxford: Blackwell, 1996. Print.
Gabrielatos, Costas, and Anna Marchi. “Keyness: Matching Metrics to Definitions.” 5 Nov. 2011.

Lancaster EPrints. Web. 10 July 2013.

Holmes, David I. “Vocabulary Richness and the Book of Mormon: A Stylometric Analysis of Mormon Scripture.” Research in Humanities Computing 3: Selected Papers from the ALLC/ACH Conference, Tempe, Arizona, March 1991. Eds. Don Ross and Dan Brink. Oxford: Clarendon, 1994. 18-31. Print.

Irwin, William. Intentionalist Interpretation: A Philosophical Explanation and Defense. Westport: Greenwood, 1999. Print.

Jespersen, Otto. A Modern English Grammar on Historical Principles. London: George Allen and Unwin, 1942. Print.

Juhl, P. D. Interpretation: An Essay in the Philosophy of Literary Criticism. Princeton: Princeton UP, 1986. Print.

Sabol, C. Ruth. “Focus and Attribution in Ford and Conrad: The Attributive Relative That in The Good Soldier and Lord Jim.” Computing in the Humanities. Ed. Richard W. Bailey. Amsterdam: North-Holland, 1982. 47-58. Print.

The Dual Materiality of Typewriters in Digital Culture

Ya-Ju Yeh

The typewriter is considered to be one of the greatest inventions that launched the modern era of writing technology. Developing through continual tinkering in the late 1860s, it became an indispensable tool used by professional writers in private homes or for business correspondence in offices. From cumbersome to portable, from noisy to noiseless, and from manual to electric, the typewriter remained a popular technological commodity for decades before personal computers came into existence. By the end of the 1980s, personal computers largely displaced typewriters, and the typewriter has almost entirely disappeared from the office. Most typewriters were relegated to museum as artifacts or to individual collections for preservation or display.

The typewriter seems to have lost its practical value as well as socio-technological significance for general users. However, its authentic, tangible materiality—the distinctive sound of click and keystroke, the unique design of keys and keyboard, and the highly recognizable pattern of slender, erect typed script—make contemporary artists, musicians, and culture curators enthusiastic to employ it in their own creative acts and works. These acts and works regarding the typewriter are re-produced in virtual representations thanks to the digital technology that emerged in the late 1990s and early 2000s. The digital representation of ideas and concepts of the typewriter, in short, its digital materiality, exerts more profound influence upon the new generation's knowledge of the typewriter. The ostensibly out-of-date typewriter revives and presents itself as a new subject matter derived from digital technology. The typewriter, once an ingenious technological material object in the market, is currently transformed into a digital material object with a great variety of applications.

This paper aims, through certain demonstrations of short video clips from websites and YouTube, to explore the dual materiality of the typewriter, or more specifically, the intersections of the typewriter originally as a material object and as a digital object in contemporary digital culture. There are some pivotal issues to be discussed: what are features of the typewriter in digital media? How does this digital materiality differ from its previous material existence? What are effects of such digital materiality? Addressing the above concerns, this paper delves into specific applications and representations of the typewriter: for instance, the typewriter as a percussion instrument in music, pantomime or comedy acts of typing either with or without substantial typewriters, and the typewriter effect and the typewriter font as one of the effects in the film maker software. This creation and collaboration is evidence of innovative expressions of digital objects. The typewriter in digital culture overturns its role as a writing device in the context of closure and stimulates human experiences of material objects, thus rendering an ideal exemplar to inflect and inform our future encounters with digital objects, and unearthing how the learners of the new generation will participate in the world of typewriters and typewriting in the twenty-first century.

Keywords: materiality, typewriter, object, digital material, digital culture

Visual Representation of the Body and East Asian Modernity: – Some Fundamentals of Digital Visual Archive and Cultural Map for the Corporeal Language in the Modern Era

Sung-do Kim and Minhyoung Kim

The primary objective of this research is to construct a digital visual archive of the corporeal language in the modern era of East Asia and to create its cultural map. Of primary importance is the challenge to enhance a deeper understanding of the East Asian modernity by corporeal representations through a variety of media. Therefore, this study focuses on the transitional period from traditional to modern society and analyzes a wide collection of visual materials, and in so doing we aim for increased comprehension the modalities of body images and its spatio-temporal implication, as derived from the modern society and culture of East Asia, and also to establish theoretical and methodological foundations for the digital humanities of East Asia.

Viewing the body as a semiotic anchor and catalyst, this study is an attempt to obtain greater understanding of the modern conceptualization of the body as an important milestone in the history of visual representation of the modern East Asia. More specifically, this work is fundamental research designed to constitute a genealogy of the modern body through the construction of visual archive and cultural map for the corporeal language in the modern era of East Asia. For a more comprehensive analysis, we have applied a good amount of time and energy to initiate a typology of various spaces where the corporeal representations have emerged, the details of which will be demonstrated later in this presentation. By setting up fourteen spatial domains of corporeality to categorize where the body operates as an agent of action, including labor, amusement, ritual, transferring, discipline, consuming, and so on, we would thus like to present the body images of the modern East Asia as a motive force, which demonstrates various imaginary stages and cultural practices within the region. As the French sociologist Bourdieu remarks, the corporeal representations are signs and symbols of “the distinction” (1984), which implicitly or explicitly express the values of contemporary age. With the influence by Western civilization, the body of modern East Asia has become a place of personal preference and individual differences, that is, a place of distinction beyond conventions.

In addition, we would like to disclose sub-topics such as the birth of the personal body, the invention of facial expressions, the body as an agent of individualization, the body imagery of aging, bio-politics in the modern visual regime, and so on. The expressions and representations of the body are profoundly determined by two elements: 1) media technologies, and 2) ideology including social norms, cultural practices, and ethical beliefs. The body is a strong semiotic medium that reflects and shapes differences and identities, and analyzes visual expressions of the body so that it leads us to discuss numerous issues of ritual, violence, family, military and public place. Our key questions are as follows: What is the place and function of the body in the modernity of East Asia?; How can we go beyond a formal and micro semiotic analysis and reconstruct the genealogies of the corporeal representations through an analysis of different visual media?; What impact has the introduction of modern visual technologies had on the expression and representation of the body in the modern time of East Asia?

In conclusion, this study attempts to grasp the relationship between modernity and corporeality in East Asia. The corporeal language especially includes body, face, and gesture, under the tradition of individual culture and the dynamics of social interaction, so that it reflects inherent characteristics of each particular time and space. Our research, therefore, intends to present a foundation, which empirically accumulates multiple corporeal languages and comparatively builds a cultural map throughout those regions of China, Japan, and Korea, as they have shared specific experiences of modernization and its dynamics. Furthermore, by supplying source references of corporeal language representation, as related to the media industry, we would like to contribute to the comprehensive management of visual informatics and its applications.

Keywords: body, visual representation, digital visual archive, cultural map, East Asian modernity

The Julfa cemetery digital repatrition project: countering cultural genocide through technology

Judith Crispin and Harold Short

The roots of Armenian culture can be traced to the establishment of Nakhichevan during 3669 BC in what is now Azerbaijan. Nakhichevan’s name derives from the Armenian “Nakhnakan Ichevan" Նախնական Իջևան (landing place), referring to the place Noah landed his Ark after the biblical deluge. It was in Nakichevan that Mesrob Mashtots first created the Armenian Alphabet and opened early Armenian schools. The centre of Nakichevan’s culture was the ancient city of Julfa (or Jugha), destroyed by order of Shah Abbas in 1605. Remarkably, Shah Abbas, recognising the extraordinary beauty and significance of Julfa’s cemetery, ordered his soldiers to leave it untouched.

Until 2005, Julfa cemetery graced the banks of the river Arax with 10,000 ornate Armenian khachkars (cross-stones) from the 15th and 16th century, inscribed with Christian crosses, suns, flowers and climbing plants. Alongside Julfa’s khachkars stood heavily inscribed ram-shaped stones, unique to this cemetery, and ordinary tombstones. Spread over three hills on Nakhichevan’s border with Iran, Julfa cemetery was home to the largest collection of East Christian cultural monuments on earth.

In 2005, in direct violation of the 1948 UN Convention on Cultural Heritage, Azerbaijani authorities demolished Julfa cemetery’s priceless khachkars with bulldozers, loaded the crushed fragments onto trucks and emptied them into the river Arax. Video footage and photographs taken from the Iranian bank of the river show almost 100 Azerbaijani servicemen destroying Julfa’s khachkars with sledgehammers and other tools. Demands by The European Parliament in 2006 that Azerbaijan “allow a European Parliament delegation to visit the archaeological site at Julfa”, were refused. Shortly thereafter, Nakhichevan authorities constructed a military shooting range on the very ground where thousands of human remains lie, now unmarked.

In 2013 we assembled a research team and embarked upon a series of pilot projects to ascertain whether sufficient primary sources still existed to make possible a large scale digital recreation of Julfa cemetery. The results were published in the ebook “Recovering a Lost Armenian Cemetery”, which can be downloaded from our webpage at (https://julfaproject.wordpress.com/). Having completed this pilot work, we have now launched a large-scale research project. The Julfa Cemetery Digital Repatriation Project aims to return to the Armenian people the entire medieval section of Julfa’s cemetery, consisting of 2000 khachkars and ram-shaped stones now designated by UNESCO as ‘intangible world heritage’.

Project outcomes will include at least two permanent 3D installations, in Yerevan and Sydney, and an international touring installation. Further outcomes will include an online virtual reality version of the cemetery installation; a permanent archive of historical photographs, facsimile illustrated manuscripts, maps, journals and other documents, which will be available online, and in the State Library of New South Wales and the Australian Catholic University. We are working with cutting edge 3D The world can be a dangerous place for monuments and it is clear that similar restoration projects will be needed to address recent high profile cultural losses in the Middle and Near East. To assist such projects, and in the spirit of collegiality, we are actively seeking partnerships and will continuously upload project outcomes and methodologies to our website. In this way we hope to develop best practice approaches to countering cultural genocide by digital recreation and repatriation of destroyed monuments.

The differences of connotations between two flowers, plum and cherry, in classical Japanese poetry, 10th century.

Hilofumi Yamamoto

Objectives:
This project will address an analysis of connotations of flowers in classical Japanese poetry: i. e., ‘ume’ (plum) and ‘sakura’. we will relatively identify the characteristics of two birds by computer modeling. Using parallel texts of original texts and contemporary translations of classical Japanese poetry, the Kokinshū, we will clarify the details of language change within Japanese in an objective procedural manner that is not influenced by human observations.

Problem:
Many scholars of classical Japanese poetry have tried to explain the constructions of poetic vocabulary using their intuition or the experiences they accumulated during their studies. Thus, they often produce modern Japanese translations through means of holistic explanations of each poem since they, even as specialist in classical Japanese poetry, cannot adequately explain the precise meanings of all words. Even if they could clearly explain all the words, their translations could include contradictions with their own explanations. This shows that translations include possibilities of non-literal elements, which are not expressible using ordinary word explanations.

To find the non-literal elements, we cannot manually match classical words in poem texts with modern words in translation texts. The problem of manual matchings of word for word from a parallel corpus, especially one comprised of classical texts, is that the act of judgment in identifying correspondences can lead to a loss of the original meaning of a word, since our present knowledge of classical words is conjectured and classified based on our knowledge of modern language.
To this end, it is necessary to employ computer assisted correspondence methods without relying on this human knowledge. I therefore use the network analysis to estimate corresponding pairs of classical Japanese words and their modern Japanese translations.

Methods:
I constructed network models of two flowers, ume (plum) and sakura (cherry), with both original poem (OP) data and contemporary translation (CT) data. I proposed the method of co-occurrence weight for digesting the huge number of co-occurrence patterns, and evaluated the methods of the modelling. Based on the network models, I will analyse the differences between ume (plum) and sakura (cherry) by subtracting the classical elements of original texts from the contemporary elements in translations. As a residual of the subtracted elements, I will observe the non-literal elements connotated in the original texts.

Materials:
I will use the Kokinshū with ten corresponding sets of modern translations. The Kokinshū is an anthology compiled under imperial orders (ca. 905). The Kokinshū consists of 1,111 poems including long poems (chōka) and head-repeating poems (sedōka) which are not short poems (tanka; 5/7/5/7/7 syllable style). I will only use the short poem form, which amounts to 1,000 poems, for stylistic consistency. The ten sets of modern translations of the Kokinshū are translated from 1927 to 1998 by ten Japanese poetry scholars.

Results:

I found that the association of poetic vocabulary is established by not only common nouns but also by proper nouns. I observed proper nouns such as place names, Kurabu, Tatsuta, Otowa, Yoshino in the network models of common nouns, and concluded that they seem to strongly influence the associations of poetic vocabulary.

The distinctions among proper nouns appearing in classical Japanese poetry has not yet been examined thoroughly in this project. Therefore, it will be necessary to examine not only common nouns but also the distinctions of proper nouns in order to further examine the connotative associations of poetic vocabulary.

The relative salience clearly indicates that both ume (plum) and sakura (cherry) share Kurabu yama (Mt. Kurabu), which comprises a cluster of nodes in the sub-network.

Quantitative Analysis of Traditional Japanese Folk Songs in Kyushu Region

Akihiro Kawase

Introduction

The ultimate goal of this study is to effectively foster and quicken the research in ethnomusicology from a digital humanities point of view by constructing a database of Japanese folk songs, which makes it possible to grasp and search for characteristics of songs with their regional locations or ages. The main purpose of this study is to partially materialize this concept by estimating the pitch range by extracting the pitch transition patterns from pieces of traditional Japanese folk songs, and by making comparisons with Fumio Koizumi's tetrachord theory in order to make the regional classification by the tendency in pitch information.

Tetrachord thoery

The Japanese musicologist Fumio Koizumi conceived and developed his tetrachord theory in order to account for traditional Japanese music. The tetrachord is a unit consisting of two stable outlining tones with the interval of a perfect fourth pitch, and one unstable intermediate tone located between them. Depending on the position of the intermediate tone, four different types of tetrachords can be formed (see Table1).

Table1: Four basic types of tetrachords

Type	Name	Pitch intervals
1	minyo	minor 3rd+major 2nd
2	miyakobushi	minor 2nd+major 3rd
3	ritsu	major 2nd+minor 3rd
4	ryukyu	major 3rd+minor 2nd

In order to quantitatively extract pitch transition patterns from Japanese folk songs, we sample and digitize folk songs included in the "Nihon minyo taikan", which is a collection of catalogued texts for Japanese folk songs. We sampled all songs from each of the seven prefectures in the Kyushu region (located in the southern part of Japan’s main island). In total, there were 1,010 songs for seven prefectures (Table2).

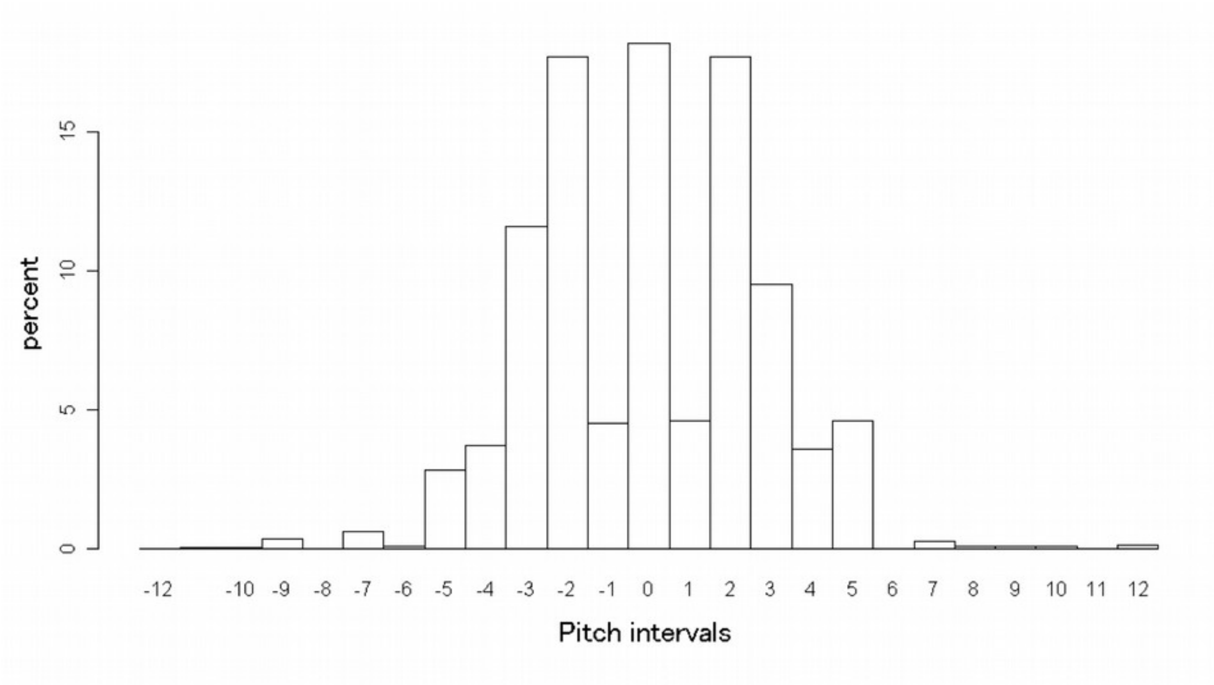
Table2: Number of songs for each prefecture

Prefecture	Songs
Fukuoka	154
Saga	81
Nagasaki	133
Kumamoto	166
Oita	130
Miyazaki	121
Kagoshima	225

Total	1,010

Procedure
In order to digitize the Japanese folk song pieces, we generate a sequence of notes by converting the music file into MusicXML file format. We devised a method of digitizing each note in terms of its relative pitch by subtracting the next pitch (height) for a given MusicXML. It is possible to generate a sequence that carries information about the pitch interval to the next note. We treat this sequence as a categorical time series, and execute n-gram analysis to measure its major patterns.
Results As a result, the findings that the first transitions of pitch intervals accounted for taking a value between +5 and -5 (within the limits of perfect fourth pitch in both descending and ascending order) (see Figure 1) suggest that most patterns consist of bi-gram or tri-gram tend to form a perfect forth pitch or return to the initial pitch at the start of the transition note. Moreover, slight differences among seven prefectures are confirmed by comparing the probabilities for Koizumi's tetrachords.

Figure 1: First transition frequency



Conclusions
In order to estimate the pitch range and pitch transition patterns, this study focused on the melodies of endangered Japanese folks songs, which were created by anonymous nonprofessional musicians and have been orally handed down from primitive times. The results indicate that pitch transition patterns of Japanese folk songs have an overwhelming tendency to either form perfect fourth pitches or return to initial pitch within two or three steps, which meets our expectation drawn from our previous study. While the intermediate tone of the tetrarchord turned out to be the salient characteristic by which to differentiate the melodies of each prefecture, it alone is not sufficient to provide a powerful explanation from which to consider differences within diverse aspects of traditional Japanese folk songs, such as relationship between linguistics and musical structure. In further research, we will construct a database holding information on both melodies and lyrics of all folk songs catalogued in “Nihon minyo taikan” in order for researchers to analyze Japanese folk songs from a digital humanities point of view.

Acknowledgments This work was mainly supported by the Japanese Society for the Promotion of Science (JSPS) Grants-in-aid for Scientific Research (15K21601) and the Suntory Foundation Research Grants for Young Scholars, 2014.

References

-Koizumi, F. (1958) Studies on Traditional Music of Japan 1, Ongaku no tomosha.
-Kawase, A. and Tokosumi, A. (2011) Regional classification of traditional Japanese folk songs – classification using cluster analysis, International Journal of Affective Engineering 10 (1): 19-27.
-Kawase, A. (2014) Structural extraction from large music corpora of Japanese folk songs, NINJAL Research Paper 7: 121-150.
-Nihon Hoso Kyokai (1944-1993) Nihon Minyo Taikan (Anthology of Japanese Folk Songs).
-MusicXML, http://www.musicxml.com/for-developers/ [accessed 10 May 2015]

Quantitative Analysis of Dissonance in Solo Piano Works: Extracting Characteristics of Debussy and Bach

Aya Kanzawa, Akihiro Kawase, Hajime Murai and Takehiro Inohara

The purpose of this study is to highlight how modern classical music came to employ dissonance, which had typically been used less often in early classical music. We analyzed musical works by C. A. Debussy, who is said to use dissonance frequently in classical music, and compared these works with those of J. S. Bach, who represents early classical music.

Today, it is generally agreed that dissonance came to be used frequently in classical music from the latter half of the nineteenth century onwards. Around this time, the collapse of tonal music, which lasted for about 300 years from the Baroque School to the Romantic School, gave way to modern music. Moreover, Debussy is often seen to have introduced a new era. Debussy was not caught up in the existing concept of functional harmony, and formed impressionist music by using a unique harmony including unresolved dissonance [1, 2]. According to the definition of musical grammar, intervals are classified as consonant or dissonant according to the degree of consonance between two sounds when they are played at the same time [3]. Impressionist music is characterized by heavy use of dissonance. Regarding the history of harmony, including dissonance, classical music has been analyzed as a case study using the humanities method [4, 5, 6]. However, it lacks objectivity and empirical analysis, since large-scale music groups have not been targeted. Few demonstrative analyses of music have been conducted to clarify why complicated harmony—which was generally not accepted in classical harmony theory—became pervasive during modern times and came to be accepted as a comfortable sound [7].

In this study, sixty-nine solo piano works by Debussy and ten solo piano works sampled from The Well-Tempered Clavier by Bach were targeted. We consequently analyzed the musical scores of the seventy-nine above-mentioned works. Solo piano works were particularly useful because they included only one instrument, which minimized noise, and they provided limited variables. Musical scores are more suitable for analyzing than recorded sounds, allowing for comparison among composers from various times, as many musical scores written during that time used unified rules. The specific steps of analysis are as follows: first, we built a music corpus by collecting Debussy’s solo piano pieces as data. Secondly, we extracted any sounds that can be heard simultaneously from MusicXML. Thirdly, we determined each interval by noting the combination of the two sounds being heard simultaneously, and then classified them as consonance or dissonance based on the definition. Finally, we extracted statistics regarding the frequency of the use of dissonance, its density in music, and any repeated patterns.

As a result, it is evident that the frequency of dissonance in Debussy’s works is greater than in Bach’s works. However, Debussy did not often use dissonance in his early works. The frequency of dissonance in Debussy’s works increases during the early to mid period of his life and decreases during the mid to late years of his life. These results indicate that Debussy’s early style was akin to that of classical tonal music, and the composer gradually established his new style, utilizing dissonance for the first time.

In the future, we will expand the corpus size by collecting other composer’s pieces as data. We will clarify how modern classical music came to employ dissonance.

References

[1] Minao Shibata, The New Grove Dictionary of Music and Musicians, Tokyo: Kodansha, 1993.
[2] Natsuki Maeda, “A Study of Harmonic Beauty and Symbolism of Debussy—Along with an Analysis of ‘Clair de lune from Suite bergamasque’—,” Mejiro Journal of Humanities, 9, pp. 273−282, 2013.
[3] Mareo Ishiketa, Gakuten: Riron to Jisshu, Tokyo: Ongakunotomosha, 2012.
[4] Tagiru Fujii, “Auf der Suche nach der verlorenen Musik (3),” Studies in Language and Culture, 22(2), pp. 173-187, 2001.
[5] Akio Nishizawa, “On the Harmony of Debussy—The Development of His Cadence—,” Educational Bulletin of Yokohama National University, 12, pp. 110-123, 1972.
[6] Takashi Nakayama, “An Analysis of Harmony of F. Chopin’s Complete Works: Polonaisen,” Bulletin of the Faculty of Education, Kumamoto University, 46, pp. 77-100, 1997.
[7] Dmitri Tymoczko, “The Geometry of Musical Chords,” Science, 313(5783), pp. 72-74, 2006.

Image Processing for Historical Manuscripts: Character Search on Chinese Handwriting Images

Yun-Cheng Tsai, Hou-Ieong Ho and Shih-Pei Chen

According to Antonacopoulos and Downton (2007), the process of handwritten documents recognition is an emerging research topic. Dineshkumar and Suganthi (2013) mentioned that handwritten documents recognition still has many challenges to researchers. Because the human transcription process remains tedious and expensive, lots of ancient materials have been digitized as image, but yet not in full text form. Although online handwriting recognition methods can handle a variety of writing styles including cursive script since they have access to the order of strokes when a character is written. There is no access to the order of strokes as an important features for recognizing historical handwritten documents. On the other hand, optical character recognition (OCR), while performing well in recognizing printed texts, cannot fully recognize handwriting characters. It results in that users cannot search the content in digitized images of handwritten. In Figure 1, we are showing an OCR result of a Chinese handwritten text, the original image of which is shown in Figure 2. Therefore, recognizing words and characters from historical manuscripts remains an unsolved problem.

Our work focus on handwritten Chinese documents recognition. Our experimental data are the digital images of Shenxian Zhuan of the Siku Quanshu edition available on the Chinese Text Project (http://ctext.org/). See Figures 3 and 4. Approaches dealing with western languages (see Hu et al. (1996); Pradeep et al. (2011)) have been proposed with good results. However, they rely on features of alphabets which cannot be applied to Chinese. In this paper, we propose a segmentation workflow and an image similarity algorithm for offline handwritten Kanji or Chinese recognition. Our goal is to allow users to run full-text search on images of manuscripts. When a user decides on a handwritten character image as the search target, our recognition algorithm can find where the same characters appear in the same and other manuscript images. This makes characters search in digitized manuscript image possible. For example, Figure 5 shows an image from Shexian Zhuan as our input material and the caracter Huang, framed by the user in blue, as the search target. The goal is to find all the occurrences of the same character in 20 pages of images of the experiment set. The search results found by our algorithm will be framed in red, as shown in Figures 6 and 7. Our method contains two steps. The first step is a segmentation workflow to isolate the characters in the given images. The second step is an image similarity algorithm to compare each segmented character image against the search target and to determine if they are the same character.

In our segmentation workflow part, we apply 2-dimensional median filter and grey-level normalization to reduce background noise of the input material (see Pei and Lin (1995); Chun-Yu et al. (2009)). Then, we use the Hough transform to detect lines of historical handwriting manuscripts (see Ballard (1981)). Next, we use the connected-component labeling to group an area of a word (see Samet and Tamminen (1988)) in order to segment the texts into lines. In Figure 8, our algorithm segments every page of the given manuscripts images into individual lines automatically. After the line segmentation, we use sliding-window approach to segment individual characters (see Chu (1995)), as presented in Figure 9. Note that at this step not all the segmented character images are correct. But the noises don’t matter since they will not be matched when being compared against a given target character. The next step is the image similarity algorithm.

As an example, the target character image is presented in Figure 5. In order to compare all the character images of the given manuscript with the target character image, we need to find the gravity center of target character image and align all character images of manuscripts to the same gravity center (see Li and Chutatape (2000)). Then, we use the Harris corner detection to calculate corner metric matrix and find corners in character images (see Harris and Stephens (1988)). In Figure 10, the blue circle is the gravity center of character image and the green crosses are the corners of images. We sort the corners into a logical order and give them numbers to indicate the order as shown in Figure 11. Based on Danielsson (1980), our algorithm sorts the distance between every corner and grid origin. We then define our similarity function which includes the number of corners and corner distance.

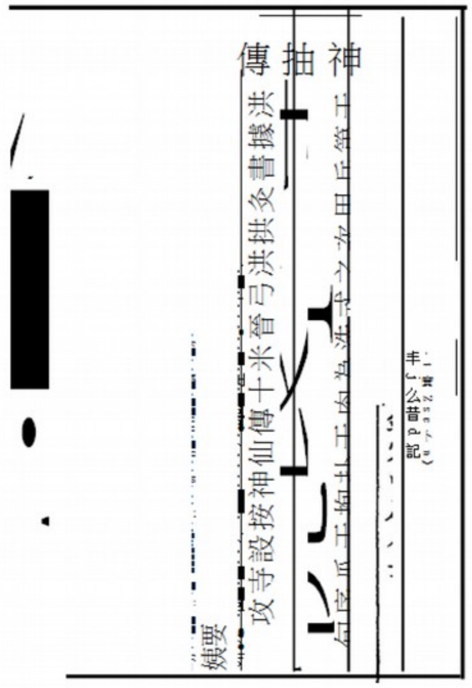
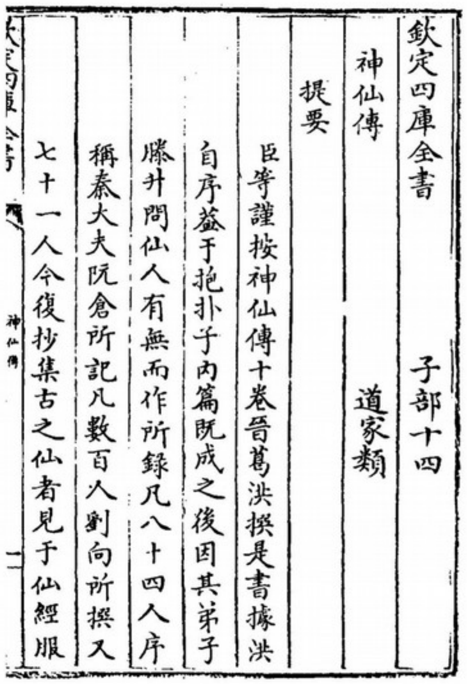


Figure 1: OCR results of Siku Quanshu.



欽定四庫全書　子部十四

神仙傳　道家類

提要

臣等謹按神仙傳十卷晉葛洪撰是書據洪

自序蓋于抱朴子內篇既成之後因其弟子

滕升問仙人有無而作所錄凡八十四人序

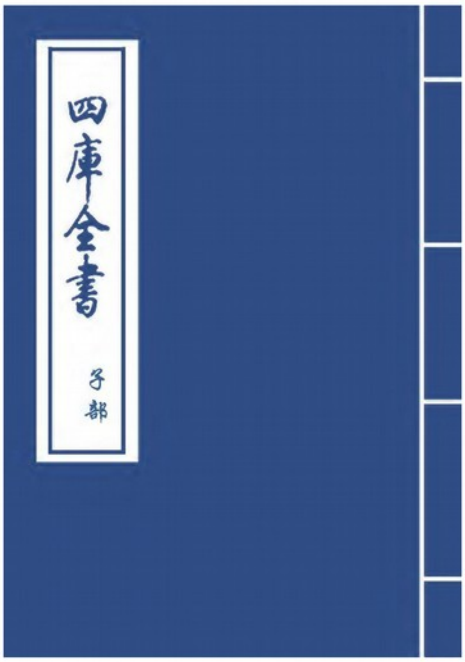
稱秦大夫阮倉所記凡數百人劉向所撰又

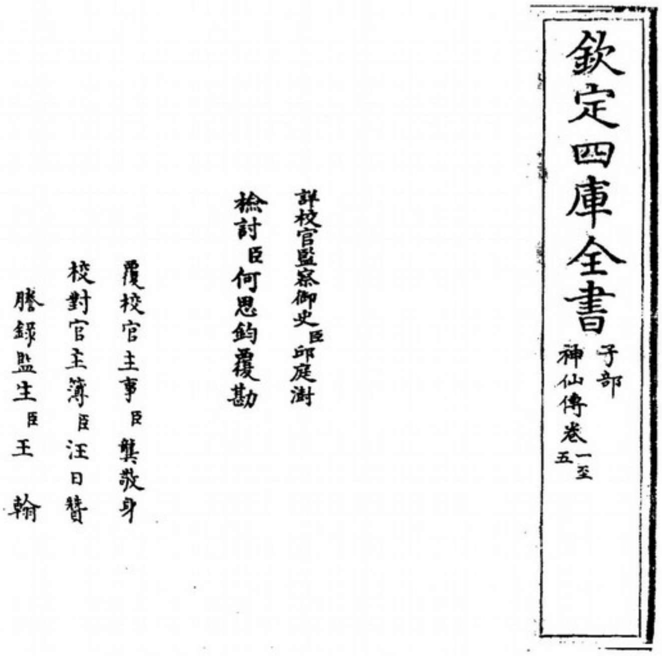
七十一人今復抄集古之仙者見于仙經服

Figure 2: One page content of Siku Quanshu.

The similarity function calculates corner distance between every character image and the target image. Based on the similarity function, the same character written in the text should have similar corners and distance in the character image. We show three examples in Figure 12. The search target is presented on the top left corner, and the top right image is one of the correct search result. The other two characters are nonmatches, meaning they did not pass the similarity function. We run the experiment with 20 pages from Shenxian Zhuan and successfully found all the occurrences of Huang among all the segmented character images. The precision is 100%. The accuracy rate of our segmentation algorith, is around 85%, which does not have great impact on our second step when comparing character images for searching.

Keywords: Handwritten Recognition, Sliding-window Approach, Grey-level Normalization, Hough Transforms, Connected-component Labeling, Harris Corner Detection.




欽定四庫全書 子部 神仙傳卷一至五

詳校官監察御史臣邱庭澍

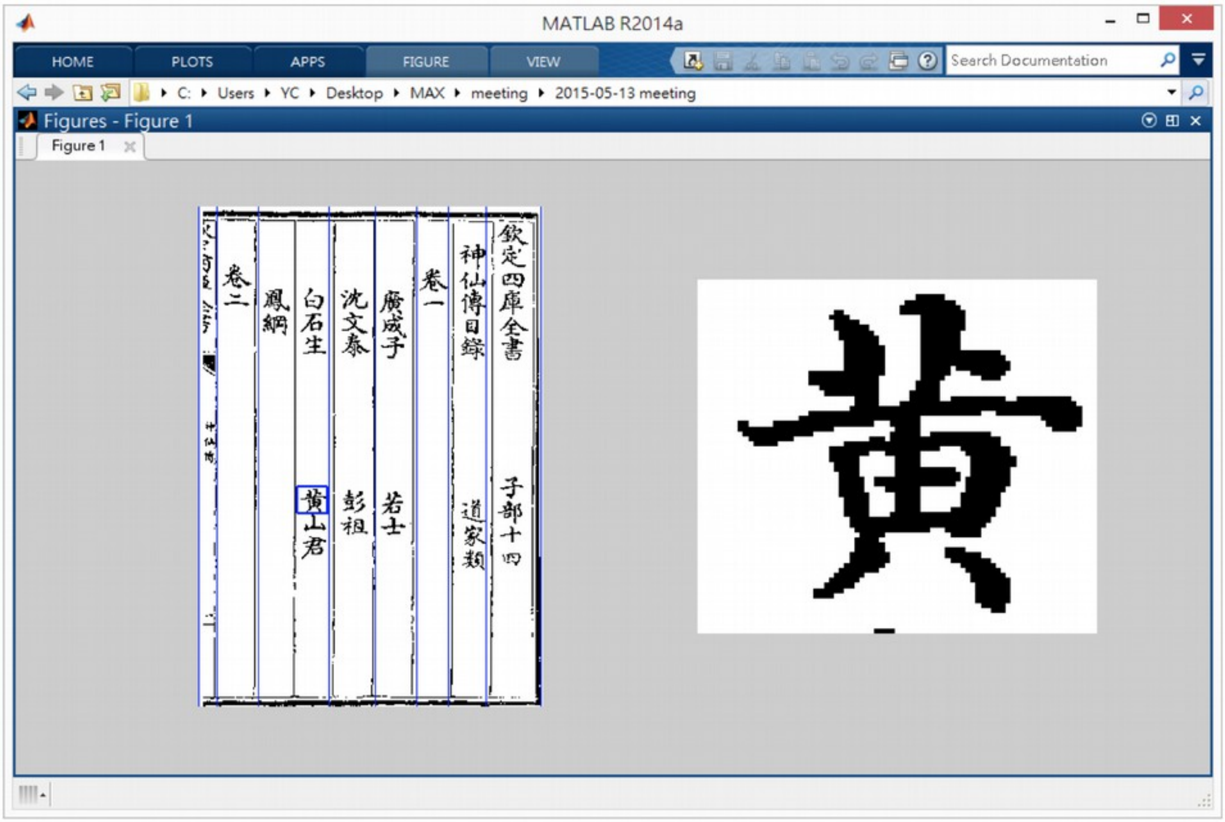
檢討臣何思鈞覆勘

覆校官主事臣龔敬身

校對官主簿臣汪日贊

謄錄監生臣王翰

Figure 3: The book cover of Siku Quanshu. Figure 4: The first page of Siku Quanshu.



欽定四庫全書 子部十四

神仙傳目錄 道家類

卷一

廣成子 若士

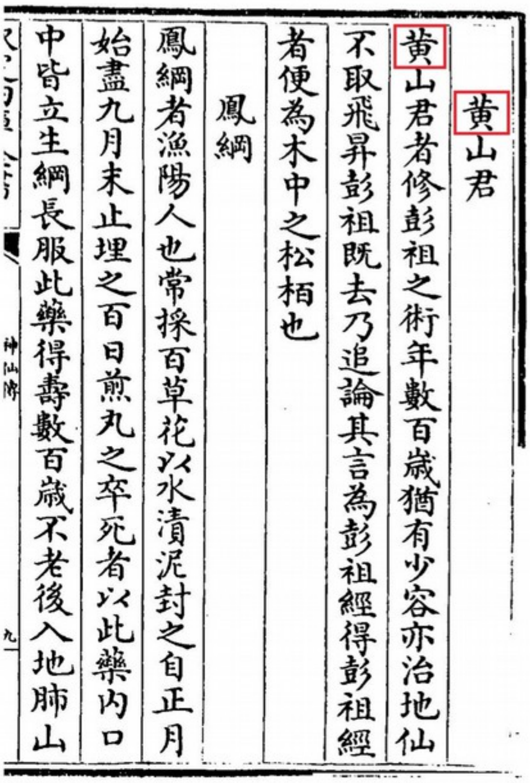
沈文泰 彭祖

白石生 黃山君

鳳綱

卷二

Figure 5: The search target Huang.



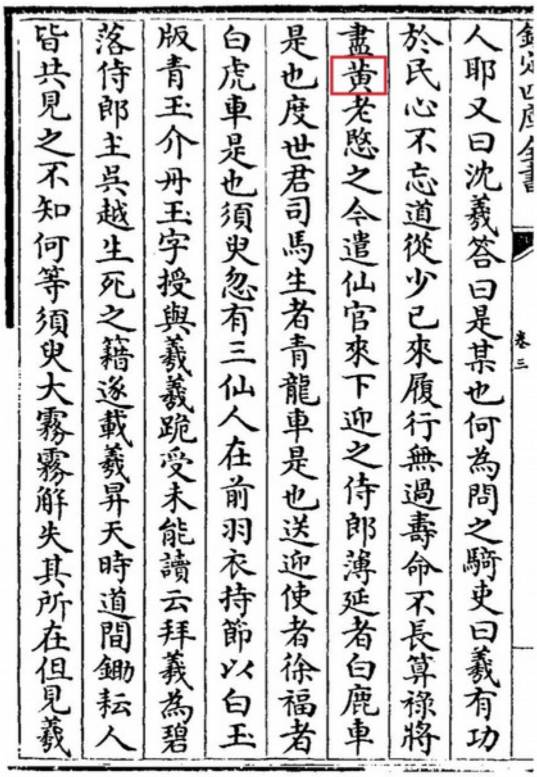
黃山君

黃山君者修彭祖之術年數百歲猶有少容亦治地仙
不取飛昇彭祖既去乃追論其言為彭祖經得彭祖經
者便為木中之松栢也

鳳綱

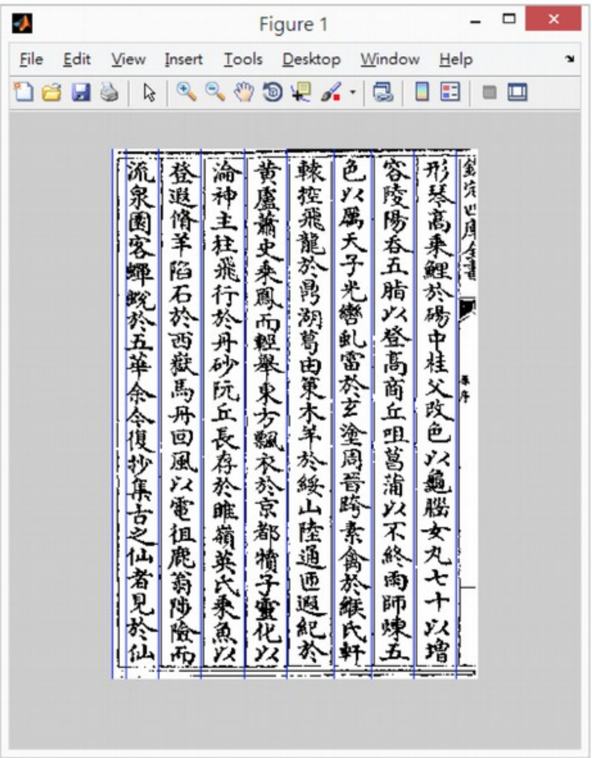
鳳綱者漁陽人也常採百草花以水漬泥封之自正月
始盡九月末止埋之百日煎丸之卒死者以此藥内口
中皆立生綱長服此藥得壽數百歲不老後入地肺山

Figure 6: The search subresult one.



人耶又曰沈羲答曰是某也何為問之騎吏曰羲有功
於民心不忘道從少已來履行無過壽命不長筭祿將
盡黃老愍之今遣仙官來下迎之侍郎薄延者白鹿車
是也度世君司馬生者青龍車是也送迎使者徐福者
白虎車是也須臾忽有三仙人在前羽衣持節以白玉
版青玉介丹玉字授與羲羲跪受未能讀云拜羲為碧
落侍郎主呉越生死之籍遂載羲昇天時道間鋤耘人
皆共見之不知何等須臾大霧霧觧失其所在但見羲

Figure 7: The search subresult two.



形琴高乘鯉於碭中桂父改色以龜腦女丸七十以增
容陵陽吞五脂以登高商丘咀菖蒲以不終雨師煉五
色以屬天子光轡虯雷於玄塗周晉跨素禽於緱氏軒
轅控飛龍於鼎湖葛由策木羊於綏山陸通匝遐紀於
黃盧蕭史乘鳳而輕舉東方飄衣於京都犢子靈化以
淪神主柱飛行於丹砂阮丘長存於睢嶺英氏乘魚以
登遐脩羊陷石於西嶽馬丹回風以電徂鹿翁陟險而
流泉園客蟬蛻於五華余今復抄集古之仙者見於仙

Figure 8: The lines detection.

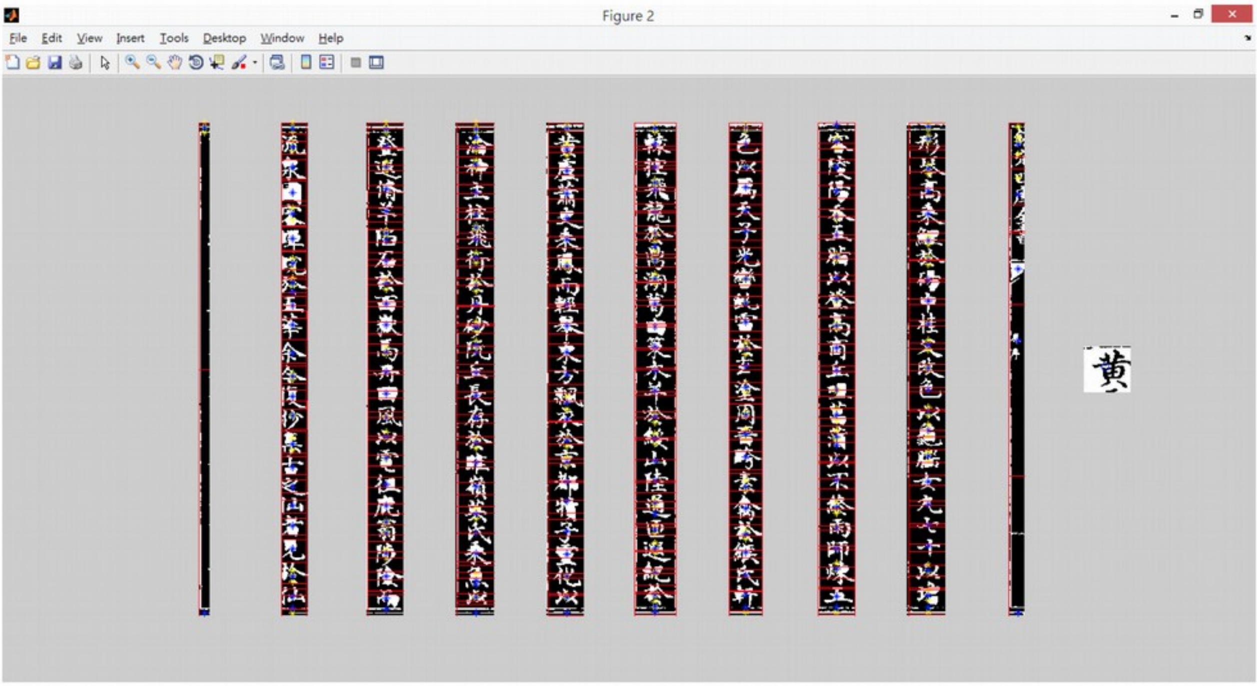


Figure 9: Separating every column into individual character.

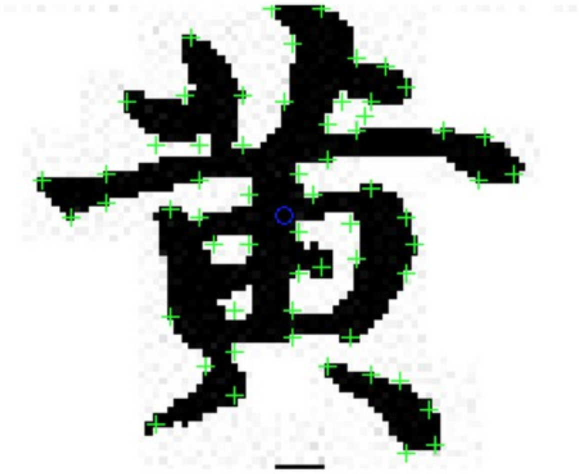


Figure 10: The corners of character image.

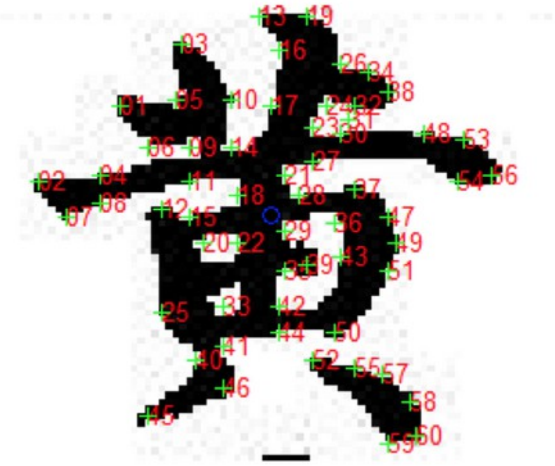


Figure 11: The corners' order.

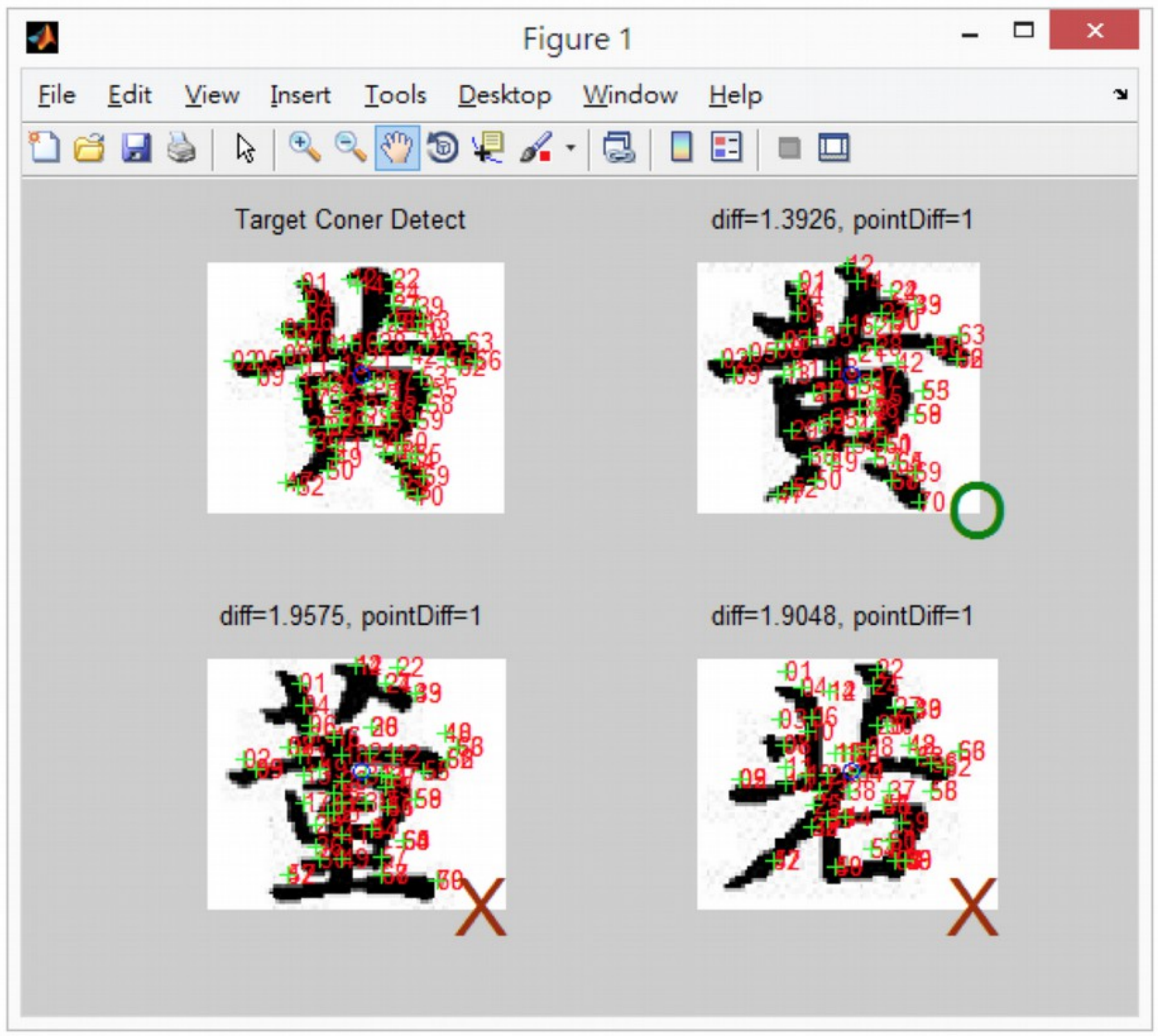


Figure 12: The three examples of similarity function. The result shows on the bottom left corner of every image. If the image passes the similarity function, the blue circle will be shown on the bottom left corner otherwise shown red cross.

References

Antonacopoulos, A. and A. C. Downton (2007). Special issue on the analysis of historical documents. International Journal on Document Analysis and Recognition 9 (2), 75–77.

Ballard, D. H. (1981). Generalizing the hough transform to detect arbitrary shapes. Pattern recognition 13 (2), 111–122.

Chu, C.-S. J. (1995). Time series segmentation: A sliding window approach. Information Sciences 85 (1), 147–173.

Chun-Yu, N., L. Shu-Fen, and Q. Ming (2009). Research on removing noise in medical image based on median filter method. In IT in Medicine & Education, 2009. ITIME'09. IEEE International Symposium on, Volume 1, pp. 384–388. IEEE.

Danielsson, P.-E. (1980). Euclidean distance mapping. Computer Graphics and image processing 14 (3), 227–248.

Dineshkumar, R. and J. Suganthi (2013). A research survey on sanskrit offline handwritten character recognition. International Journal of Scientific and Research Publications (IJSRP) 3, 13–

16.
Harris, C. and M. Stephens (1988). A combined corner and edge detector. In Alvey vision conference, Volume 15, pp. 50. Manchester, UK.
Hu, J., M. K. Brown, and W. Turin (1996). Hmm based online handwriting recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on 18 (10), 1039–1045.
Li, H. and O. Chutatape (2000). Fundus image features extraction. In Engineering in Medicine and Biology Society, 2000. Proceedings of the 22nd Annual International Conference of the IEEE, Volume 4, pp. 3071–3073. IEEE.
Pei, S.-C. and C.-N. Lin (1995). Image normalization for pattern recognition. Image and Vision computing 13 (10), 711–723.
Pradeep, J., E. Srinivasan, and S. Himavathi (2011). Diagonal based feature extraction for handwritten alphabets recognition system using neural network. arXiv preprint arXiv:1103.0365 .
Samet, H. and M. Tamminen (1988). Efficient component labeling of images of arbitrary dimension represented by linear bintrees. Pattern Analysis and Machine Intelligence, IEEE Transactions on 10 (4), 579–586.

Development of a Data-visualization Tool for Ukiyo-e Analysis: A Case Study of Otohime

Shinya Saito and Keiko Suzuki

■ Background of this research

It is assumed that approximately a million ukiyo-e (Japanese woodblock prints) still exist all over the world. The prints are famous not only as representative popular culture of the Edo period but also because of their significant influence on European art, especially Impressionism. In order to conduct comprehensive analyses of ukiyo-e, researchers in the Humanities often need to handle a vast number of prints whose themes come under their research purview, organize and compare them to identify similarities and differences as well as historical changes in the targeted prints.

■ Purpose of this research

This research is to support such comprehensive research of ukiyo-e by developing an original, data-visualization system. As a part of the system development, this case study focuses on the prints' visual features—how the system can help analyzing them by simulating the analytical process.

■ Method of this research

For developing the system, we take an example of pictorial theme of "Otohime," who is these days explained as a mythical princess, living in Ryūgū or the Dragon Palace at the bottom of the sea. Our investigation of the ukiyo-e research's process helps identifying what the researchers actually need and how to deal with it. That also helps us decide what kinds of design and functions of the system are crucial for the ukiyo-e research.

For comparing and analyzing Otohime's visual features thoroughly, the following 9 visual features of hers are chosen as indexes: 1. headdress; 2. hairstyle; 3. frills; 4. scarf; 5. Chinese fan; 6. Chinese-style clothes; 7. collar; and 8. apron; and 9. Urashima. Checking the indexes leads researchers to understand the degree of similarity in different ukiyo-e prints, and generate hypotheses about what kinds of factors affect specific ukiyo-e production as well as historical changes in Otohime's overall imagery.

■ System Development

In order to pursue the above-mentioned purpose, we are developing the SALOMONIS system which can load a ukiyo-e dataset encoded in JSON(JavaScript Object Notation) format. The following list shows how SALOMONIS visualizes the data.

< Entire dataset and records >

In the system, a record is visualized as a line, each of which is arranged in a radial manner. As a result, SALOMONIS visualizes the records, i.e., the entire dataset, as a circle.

<Columns>

As mentioned above, a line indicates a record, and each record includes columns that correspond to indexes. Our system visualizes columns as dots, plotted on the line.

Each color of the dots indicates different value of the column.

In the case of the Otohime dataset, it has 41 records and 9 columns. Therefore, 41 lines are arranged, each of which has 9 dots.

<Interactive function>

As each line is linked to an ukiyo-e print, when the user puts the mouse cursor on a certain line, a chart appears with its corresponding ukiyo-e's 9 visual features, mentioned on the above. Using this interactive function, the user can compare ukiyo-e in a speedy and accurate manner.

<Similarity-screening function>
Further, when the user chooses one ukiyo-e as a reference point, then pushing the "Play" button, the system starts automatically screening how the other ukiyo-e are similar or not to the chosen one. Depending upon the degree of the similarity, lines of the ukiyo-e appear highlighted with colors. For example, when one ukiyo-e shares 80 percent or more similarity with the chosen one in terms of the visual features, the line appears in orange. In the case of 60 percent or more, green; 40 percent or more, yellow; and 20 percent or more, red.

■ Discussion

The staff involved in the system development tested it, whose results demonstrated considerable efficiency in comparative analyses of the prints. The system proved its efficiency as it not only saved the time for the analyses but also suggests new ways to think about the prints. Right now, however, we are still in the developing stage, planning to test the system by applying it for more cases. That will hopefully prove its efficiency scientifically. In the future, we also hope to examine its applicability not only for ukiyo-e research but also for other researches in the Humanities.

The Rise of Islamic Reformism in the Seventeenth and Eighteenth Centuries: Scholarly Network Analysis and Visualization with Gephi

Yuri Ishida (Waseda University)

This study aims to clarify the formation of Islamic Reformism. The representative person of Islamic Reformism is Muḥammad ibn ʻAbd al-Wahhāb (d. 1792), who is the father of Kingdom of Saudi Arabia and he is known as his strict religious policy that emphasizes to follow the Qur'an and the practices of the prophet Muḥammad. Even today, his idea still has influence on Islamic extremists like Islamic State (IS). It is said that Islamic Reformism rose out of Muslim scholars' interaction in seventeenth and eighteenth centuries Arabian Peninsula, where is famous as the pilgrimage destination of Muslims as well as one of academic centers in the Islamic World. Arabian Peninsula attracts Muslim scholars from North Africa, Egypt, the Eastern Mediterranean, South Asia, Southeast Asia and other regions. These scholars established master–pupil relationships and their scholarly network spread all over the Muslim World.

Azyumardi Azra reconstructed the network on the basis of various Arabic and Persian biographical works: Khulāṣa al-athar by al-Muḥibbī (d. 1699), Fawā'id al-irtiḥāl by al-Ḥamawī (d. 1711), al-Insān al-ʻayn by Shāh Walī Allāh (d. 1762), Silk al-durar by al-Murādī (d. 1791), ʻAjā'ib al-athār by al-Jabartī (d. 1825), al-Badr al-ṭāliʻ by al-Shawkānī (d. 1839), Abjad al-ʻulūm by al-Qannūjī (d. 1889), Fahrasa by al-Kattānī (d. 1927), al-Aʻlām by al-Ziriklī (d. 1976) and so on [1]. These books record the name of eminent scholars and their educational background. It enables us to follow scholars' moving and their academic lineages.

However, these biographies' information has not been organized enough because the master–disciple relationships are too complicated to express by a simple tree diagram. For example, it is ordinary for Muslim scholars to learn one subject from several scholars with changing their residence. Depended on these situations, the hub scholar of this network has over thirty disciples. To analyze the scholarly network, this study uses Gephi [2]. This open source software is already applied to visualize historical networks and then it looks proper to start with. The approach of Digital Humanities is not popular in the field of Islamic studies and then this study becomes a pioneer.

The procedure of this study is as below. First of all, the outline of scholarly network in seventeenth and eighteenth centuries Arabian Peninsula is visualized with Gephi according to Azra. Every scholar is provided ID as Node and then he is connected to his masters by Edge. If ID_2 is a disciple of ID_1, "Source" is 1, "Target" is 2, "Type" is "Directed" and "Interaction" is "Teaching." The efficacy of Gephi will be shown by the comparison with Azra's scholarly tree diagram.

Second, detail information from the nine biographies above is added on the outline of scholarly network for further analysis. Taking al-Insan al-ʻayn by Shāh Walī Allāh for instance, this book deals only twelve scholars, who are related to the author's academic lineage. In contrast to this fewness, it gives rich information about scholars' background and reveals that the academic centrality of Egypt in the seventeenth century. Previous studies pointed that North African scholars played an important role in the formative period of Islamic Reformism by diffusion of Mālikī, one of schools of Islamic law [3]. However, the data set from Shāh Walī Allāh's al-Insan al-ʻayn tells that two of three North African scholars studied in Egypt before visiting Arabian Peninsula and scholars from other regions also learned Mālikī in Egypt. It indicates that the Mālikī center was Egypt in the seventeenth century and Arabian Peninsula replaced the role in the next century. In addition, most of scholars did not change their school of law even though they emphasized al-Muwaṭṭa' by Mālik ibn Anas, the founder of Mālikī. These facts will be supported by other biographies and may throw doubt on the past explanation.

Finally, the entire picture of the Muslim scholarly network is visualized by Gephi. Its analysis will show that the shift of the academic center and main school of Islamic law between the seventeenth century and the eighteenth century. To know what is changed and what is unchanged in these two centuries is necessary to identify the elements of Islamic Reformism. This study will give a quantified conclusion.

Keywords Islamic Reformism, Asian History, Network, Visualization, Gephi

Reference

[1] Azra, Azyumardi: The Origins of Isalmic Reformism in Southeast Asia, Sydney: Allen & Unwin; Honolulu: University of Hawai'i Press (2004).
[2] < http://gephi.github.io/ > (visited on 10/07/2015).
[3] Voll, John O.: "Hadith Scholars and Tariqahs: An Ulama Group in the 18th Century Haramayn and Their Impact in the Islamic World," Journal of Asian and African Studies, Vol. 15, Issue 3-4, pp. 264–273 (1980).

Opening up Japanese Resources for a Pan-European Common Names Webservice – 5 steps towards Linked Open Data Humanities

Eveline Wandl-Vogt, Megumi Kurobe, Chitsuko Fukushima, Thierry Declerck and Heimo Rainer

The digital media is changing the way we think, and – even more quickly – the way we work and with that access and navigate in digital surroundings and make use of data and tools [1,2].

Language is the tool mostly used to access the digital world.

What is common for standard language gets challenging when working with non-standard language such as dialectal or historical data.

In this paper the authors discuss a Globalisation of an existing knowledge resource, namely the Common Names Webservice, developed by the Naturhistorisches Museum Wien (nhm) in the framework of the European Project of OpenUp! [13] in conjunction with the Austrian Academy of Sciences, Austrian Centre for Digital Humanities on the example of the Database of Bavarian dialects in Austria (DBÖ).

The authors demonstrate a step-to-step integration of new resources in their model on the example of Japanese Resources, both, from botanical-taxonomical research [e.g. http://togodb.dbcls.jp/species_names_latin_vs_japanese], as well as etymological resources [e.g. http://www.nihonjiten.com/gogen/seibutsu_hana/] or linguistic resources, thesauri [e.g. http://www.ninjal.ac.jp/archives/goihyo/]; http://www.taishukan.co.jp/item/nihongo_thesaurus/thesaurus.html ; as well as non-scientific resources [e.g. http://www.plantsindex.com/plantsindex/demo_html/top/list_sc.htm] or Wikipedia / Dbpedia.

The authors will establish their model on the example of a well known plant in Europe and Japan, which is the Bellis perennis `daisy´.

They focus on 5 ways to improve this common service and – in doing so – several humanities fields and discuss how these improvements vice versa have a positive input on digital humanities in general:

1. Convert a dictionary data | common names collections into machine readable format and publish it in the LOD [3].
The authors exemplify an existing model for lexicographic data in the cloud, namely the model of the ontology-lexica working group. [6]
Furthermore, the model used for common names service is discussed and described.

2. Create an interactive service rather than an object [4].
The authors show, how they build an interactive webservice, dealing with linguistic and botanical data. They discuss the added value for the connected team-members and embed the results into their specific workflows. They summerize with suggestions for added value for interdisciplinary research infrastructures.

3. Enrich the service with collaborative expertise.
The authors discuss on the example of the Collaboration of the Botanists at the Naturhistorisches Museum Wien (embedded into the European Project “OpenUp! Opening Up Europes Natural History Museums for Europeana”) and the Lexicographers (embedded into the European Network for electronic Lexicography (ENeL) [11] and the European Research Infrastructure DARIAH.EU) [12] at the Austrian Academy of Sciences.
Furthermore, they point out the added value established by the collaboration ob botanists and lexicographers making use of language technology and semantic web technologies as well as linked (open) data technologies.

4. Create several paths for access and analytics [5]; keep open minded for human beings, so keep all senses and disciplines in mind to assure best ways of data reusability.
The authors give several examples for re-using of those data:

a. Europeana [7]
Based on an existing example the authors discuss added values for LOD-based webservices:
Example: Cantharellus cibarius
@ Database of Bavarian Dialects in Austria:
http://wboe.oeaw.ac.at/dboe/beleg/148799
@SKOS-Service NHM:
openup.nhm-wien.ac.at/commonNames/?query={%22type%22:%22/name/common%22,%22query%22:%22Cantharellus+cibarius%22}&format=edmSkos
@Common-Names-Service NHM:
http://openup.nhm-wien.ac.at/commonNames/references/scientificName/4054
@ Europeana:
http://www.europeana.eu/portal/record/11610/_IGMF_ETI_NETHERLANDS_360.html?start=11&query=*%3A*&startPage=1&qf=PROVIDER:OpenUp!&qf=cantharellus&rows=24#

b. Wikipedia [8]

c. Babelnet [9]

d. Agrovoc [10]

5. Connect and globalize, data, services as well as people.
On the example of a new project of a European Plant Names Thesaurus developed within the framework of the COST ENeL action, the authors discuss the possibilities of interdisciplinary collaboration.
On the example of the Common Names Service, the author discusses the vision of worldwide virtual research environments, focussed on certain issues, yet embedded into global infrastructures. They offer opportunities to join in and be part of.

Concluding,

added value of Linked Data Humanities on the example of interdisciplinary controlled Vocabularies will be presented for discussion and open feedback.

References:

[1] Carr, Nicolas "Surfen im Seichten. Was das Internet mit unserem Hirn anstellt." München

2013.
[2] Nentwich, Michael, König, René "Cyberscience 2.0. Research in the Age of Digital Social Networks. " Wien 2012.
[3] Wandl-Vogt, Eveline, Declerck, Thierry "Mapping a dialectal dictionary with Linked Open Data" In Kosem, Iztok (et al., Eds.) Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia: 460-471. <http://eki.ee/elex2013/proceedings/eLex2013_32_Wandl-Vogt+Declerck.pdf> [accessed: 15.05.2015]
[4] Tasovac, Toma "Reimagining the dictionary, or why lexicography needs Digital Humanities." Digital Humanities (2010): 1-4. <http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/pdf/ab-883.pdf> [accessed: 15.05.2015]
[5] Theron, Roberto, and Laura Fontanillo "Diachronic information visualization in historical dictionaries." Information Visualisation (2013): 1473871613495844
[6] Ontology-Lexica Working Group: http://www.w3.org/community/ontolex/ [accessed: 15.05.2015]
[7] Europeana. Think Culture: http://www.europeana.eu/ [accessed: 15.05.2015]
[8] Wikipedia: http://de.wikipedia.org/wiki/Wikipedia%3AHauptseite [accessed: 15.05.2015]
[9] Babelnet: http://babelnet.org/ [accessed: 15.05.2015]
[10] Agrovoc: http://aims.fao.org/agrovoc [accessed: 15.05.2015]
[11] COST ENeL: COST IS 1305: European network of electronic Lexicography: www.elexicography.eu [accessed: 15.05.2015].
[12] DARIAH.EU: European Research Infrastructure for the Arts and Humanities: www.dariah.eu [accessed: 15.05.2015].
[13] OpenUp! Opening Up the Natural History Heritage for Europeana: http://open-up.eu/ [accessed: 15.05.2015].

Time-series methods in language corpora analysis: – cross-correlation as an exploratory tool of multiple lexical series.

Adam Pawłowski

Time-series have proven to be an efficient tool of analysis of longitudinal data in corpus linguistics. ‘Longitudinal data’, as defined here, are frequencies of lexemes (or of other meaningful units), generated from the corpora that consist of text samples annotated with meta-information with the greatest possible precision in terms of their date of creation. This allows to discover trends and periodical oscillations of the so called ‘lexical series’, representing frequencies of selected lexemes or sets of lexemes over a long period of time. It is worth adding that lexeme frequencies are a sort of projection of the phenomena or processes that develop in the real world described in texts. This is particularly true for press texts, referring perforce to current events in the external world.

The research will be carried out on the Chronological Corpus of Polish Press – ChronoPress. ChronoPress is the first corpus fully dedicated to chronological (sequential) analyses of texts. It includes 52,000 short samples of Polish press articles published between 1945 and 1954. These are annotated with exact dates of publication as well as with other relevant characteristics such as titles, names of authors etc. Every month is represented by circa 120,000 lexemes excerpted randomly from a set of mainstream newspapers (dailies, weeklies). Such corpus size makes it possible to generate relatively long series of lexeme frequencies, covering a turbulent period in the history of Poland, a period marked by the end of the Second World War and by the consequent imposition on the Polish people of the communist, totalitarian rule.

A problem that is of particular interest here is the clustering of lexemes representing similar or dissimilar behaviour over time. It may be hypothesised that some lexemes will display similar shapes of the histogram curves (e.g. holiday and travel, crisis and unemployment etc.), while some other ones will be negatively correlated. Such lexemes can be selected using intuitive external criteria or detected in the entire lexicon without any preliminary assumptions in a sort of theory-free, exploratory approach. The goal of this study is to evaluate the efficiency of the cross-correlation function used here as a tool for detecting pairs of lexemes with similar time-series patterns (based on monthly periods). It should be noted that cross-correlation is typically associated with signal analysis, but it may be applied to the analysis of any discrete, longitudinal data.

The afore-described test will involve three stages. Firstly, a set of 100 lexemes with the highest frequencies, which are present in all the samples, will be selected. Texts from the corpus will then be lemmatized. It is noteworthy that the suggested number of lexemes to be processed is in fact arbitrary and may vary depending on the size of the available corpus. Secondly, time-series for these lexemes will be generated. Thirdly, cross-correlation coefficients will be calculated for all the pairs of the time-series thus generated and represented in the form of a matrix of similarities. Should clusters of similar time-series exist, the matrix of cross-correlation coefficients will clearly flash them out. The basic test will involve the examination of the cross-correlation with a zero lag, however, additionally, non-zero lags cross-correlations may also be verified as dependencies between lexeme frequencies shifted in time cannot be excluded.

Open Stylometric System based on Multilevel Text Analysis

Maciej Eder, Maciej Piasecki and Tomasz Walkowiak

Stylometry, or statistical analysis of writing style, is aimed to investigate text to text similarity on different linguistic levels. Originally developed to verify authorship of anonymous literary works, it was later extended and generalized to assess style differentiation, chronology, genre, author's gender etc. It relies on the assumption that authors have their unique writing habits – sometimes referred to as the 'stylistic fingerprint' – that can be pinpointed using, e.g. machine-learning approaches. Classical stylometric approaches are usually focused on very simple linguistic features that can be automatically retrieved from text documents. They include: frequencies of the most frequent words, occurrences of punctuation marks, average sentence length, average word length etc. However, since these style-markers turn out to be very effective for authorship attribution and style recognition, they are not suitable, e.g., for more semantically sensitive analysis. On theoretical grounds, multilevel features based on Natural Language Engineering (NLE) should be more efficient here.

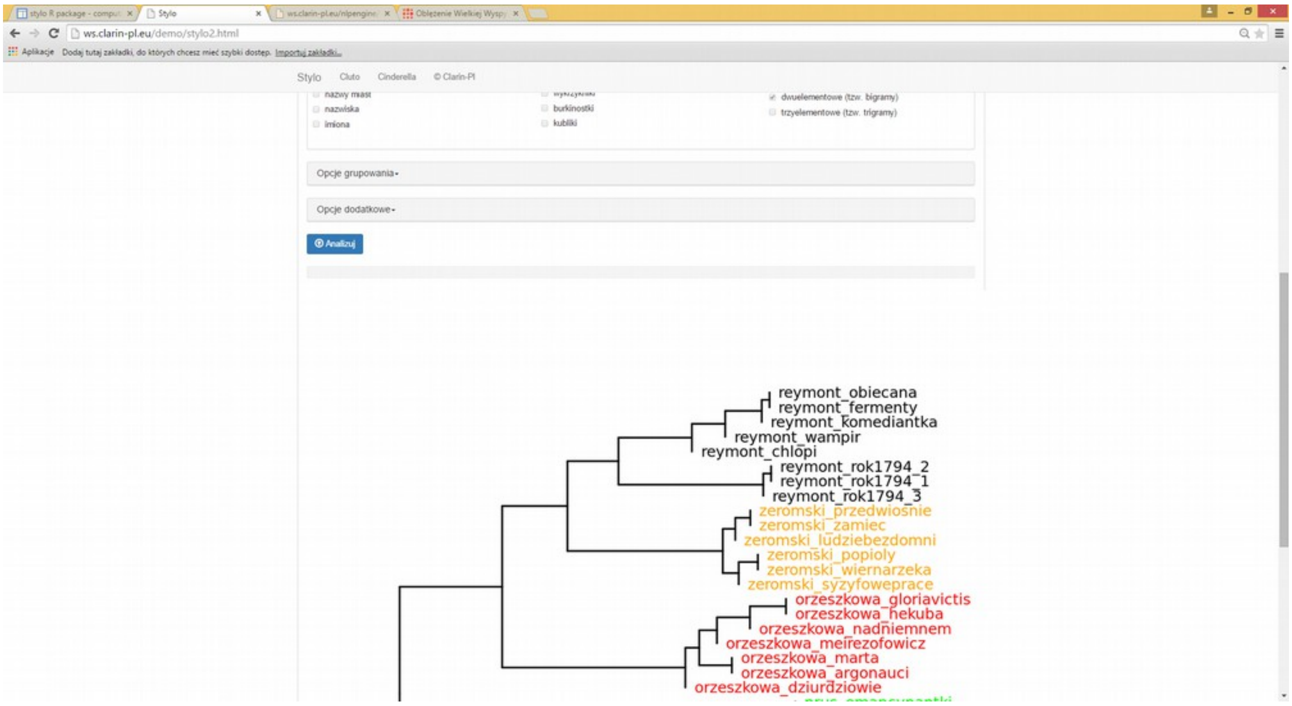
Stylometric techniques are known for their high accuracy of text classification, but at the same time they are usually quite difficult to be used by, say, an average literary scholar. Presumably, stylometry would have been routinely applied in many research tasks in the Humanities, if it had been more accessible to researchers with no programming skills. It seems that implementing some of these methods into an out-of-the-box tool might overcome the above drawback. The goal of our work was twofold, then. Firstly, we wanted to develop a web-based system for stylometry targeted at scholars in the Humanities, that does not require installing any software on local machines, and takes advantage of high-performance capabilities of the server. Secondly, we planned to enlarge the set of standard stylometric features with style-markers referring to various levels of the natural language description and based on NLE methods.

Computing word frequencies is simple in English, but in the case of highly inflected languages, characterised by a large number of possible word forms, one faces the problem of data sparseness. Thus, it might be better first to map the inflected word forms to lexemes, and next to calculate the frequencies of the lexemes. The mapping can be performed by using a morpho-syntactic tagger. The tagger tries to automatically recognise grammatical attributes of the analysed words (e.g. case, gender). Such attributes can be also used as elements of the text document description, e.g. higher frequency of the first person can signal personal style of writing. Moreover, the documents can be further processed and enriched with Proper Names identification, or even with disambiguated word senses (e.g. as recorded in a semantic lexicon). In the present paper, we will analyse the applicability of the aforementioned language tools to the document description for the needs of stylometry and semantic content-based clustering of the documents. One needs to be aware, however, that using NLE tools sometimes does not guarantee better classification precision, e.g. syntactic parsers for Polish do not improve the results as their accuracy is limited.

The workflow is as follows. Input documents are processed in parallel. Since the uploaded documents might be in different formats (doc, docx, pdf, html, rtf, txt using various codepages), they are converted to uniform text format. Next, each text is analysed by a part-of-speech tagger (we use WCRFT2 for Polish [5]) and then it is piped to a name entity recognizer (in our case it is

Liner2 [3]). When the annotation phase is completed for all the texts, the feature extraction module comes into stage (we use the tool Fextor [1]). It creates a matrix of features, which is then normalised. Finally, the R package Stylo [2] is called to perform explanatory analysis, e.g. multidimensional scaling. The results obtained in graphical format are displayed by the web browser (see Fig. 1). The web interface allows uploading input documents from a local machine or from a public repository, provides some options of selecting a feature set and options for selecting a grouping algorithm. Apart from the standard procedure, one might want to use Cluto [6], a well known clustering tool, to perform final steps of the analysis. In such a case, Cluto replaces the R package Stylo in the text processing workflow and expands the set of clustering methods that can be in used in the analysis.

The system in its current form is focused on processing Polish. English texts are analysed on the level of word forms only. However, as the stylometric procedure is conceptually language independent, we plan to extend our workflow by adding NLP tools and feature extraction techniques for other languages. Firstly, full support for English will be added, but later, the system will become ready to support any other language.



☐ nazwy miast
☐ nazwiska
☐ imiona
☐ wykrzykniki
☐ burkinostki
☐ kubliki
☑ dwuelementowe (tzw. bigramy)
☐ trzyelementowe (tzw. trigramy)

Fig.1 Stylometric system GUI

The web based interface and the lack of the technical requirements facilitates the application of text clustering methods beyond the typical tasks of the stylometry, e.g. analysis of types of blogs [4], recognition of the corpus internal structure, analysis of the subgroups and subcultures, etc.

The fully-functional system offers a variety of possible features combined with rich functionality of the clustering modules. As a result it can be used as a research tool in the stylometric analysis but also for discovering semantic classes in large document collections. Possible further developments of the system will be discussed, e.g. extraction of the descriptive features from text clusters.

Bibliography

[1] Broda, B., Kędzia, P., Marcińczuk, M., Radziszewski, A., Ramocki, R., Wardyński, A.: Fextor: A feature extraction framework for natural language processing: A case study in word sense disambiguation, relation recognition and anaphora resolution. Studies in Computational Intelligence, vol. 458, Springer, pp. 41-62 (2013)

[2] Eder, M., Kestemont, M. and Rybicki, J. (2013). Stylometry with R: a suite of tools. In: Digital Humanities 2013: Conference Abstracts. University of Nebraska--Lincoln, NE, pp. 487-89.

[3] Marcinczuk, M.; Kocon, J., Janicki, M.: Liner2 - A Customizable Framework for Proper Names Recognition for Polish. Studies in Computational Intelligence, vol. 467, pp. 231-253 (2013)

[4] Maryl Maciej. „Kim jest pisarz (w internecie?)” w: Teksty Drugie 2012 nr 6.

[5] Radziszewski, A.: A tiered CRF tagger for Polish, Intelligent Tools for Building a Scien-tific Information Platform. Studies in Computational Intelligence, vol. 467, pp. 215-230 (2013)

[6] Zhao, Ying and Karypis, George. Hierarchical Clustering Algorithms for Document Datasets. Data Mining and Knowledge Discovery, Vol. 10, No. 2, pp. 141 - 168, 2005.

Mining Rhetorical Features of Charles Dickens: A study in rhetorical profiling of style

Tomoji Tabata

This paper describes a stylometric approach to Dickens's writings, applying a state-of-the-art classification algorithm in an effort to distinguish Dickens's texts from a reference corpus of texts. The study this paper draws upon makes use of DocuScope, a text analysis environment with a suite of interactive visualisation tools for corpus-based rhetorical analysis developed at Carnegie Melon University. DocuScope makes it possible to tag texts in three rhetorical annotation layers. Any word that is tagged fits into one of 101 Language Act Types (LATs), including Abstract Concepts, Private Thinking, Comparison, Subjective Perception, Descriptive Features, etc. Any LAT fits into the "bucket" of one of 51 Dimensions (i.e. larger categories), and any Dimension fits into the larger bucket of one of 17 Rhetorical Clusters. The tagged texts can further be transformed into a frequency table profiling rhetorical features of each text in the corpus.

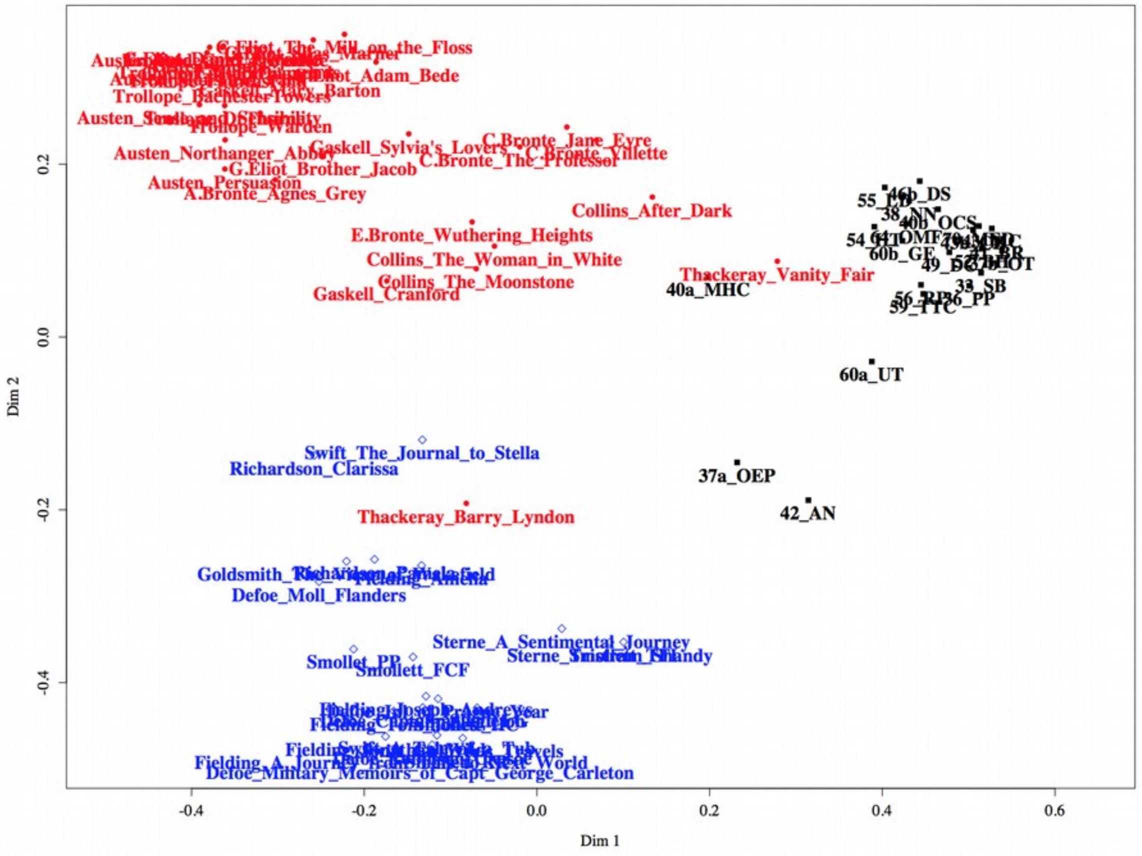


Figure 1: Random Forests analysis of Dickens corpus in comparison with 18th-and 19th-Century corpora: Dickens (black); 18th-Century texts (blue); 19th-Century texts (red)

A machine-learning classification technique (Breiman's (2001) 'Random Forests') is then used to visualise a complex interrelationships among texts in a multidimensional scaling plot (see Fig. 1) as well as to spotlight rhetorical features Dickens consistently used or avoided in his texts in comparison with the control set of texts. By demonstrating how Random Forests can be used as a powerful tool to identify Dickensian stylistic markers, this paper also proposes the feature-extraction technique detailed in this study as a more intellectually robust alternative to the traditional ‘key word’ analysis based on log-likelihood ratio scores, a popular method in corpus linguistics for extracting a set of words that characterise a particular text, a particular register/(sub)corpus, or a particular diachronic set of texts, etc. from others.

Negotiating Digital Collaboration: The Case of the Sakamaki/Hawley Collection at the University of Hawaii at Manoa Library and the University of the Ryukyus Library

Jennifer Beamer

Since 2009, the University of Hawaii at Manoa Library and the University of the Ryukyus Library have been working together to digitize and provide open access to digital resources from the Sakamaki/Hawley Collection. Presently over 218 digitized titles are now online at the University of the Ryukyus Library Ryukyu/Okinawa Special Collections Digital Archives. This digital collection is unique, as its physical contents are housed and digitized at the University of Hawaii Library, and the digital collection is archived and interfaced on servers at the University of the Ryukyus Library. This collection has many special features, including content summaries and explanations in English and Japanese, and a magnification viewing function. Modern language translations and text reprints in the original languages will debut in 2015.

While some will argue that the creation and curation of digital humanities resources is in and of itself an example of successful digital-humanities collaboration, I argue that the collaboration should move to the next level. That is to say, both libraries should investigate the opportunities for engaging in collaborative digital humanities research on the Sakamaki/Hawley items.

This paper examines the possibility of enhancing the existing technical processes (e.g., the metadata of the collections) and relational processes in an effort to expand its “digital humanities researchability”. It will propose suggestions for collaborations on enhancing the relationships between libraries when encoding various cultural resources.

First, this paper explains the technical processes for encoding of metadata of the items. If more complex metadata was encoded with the items in the digitization process, opportunities for further analysis using digital humanities tools and methods of research on the collection would present themselves. For example, TEI visualization of geographic, cultural, social or even political relationships (to name but a few themes) could be mapped and used in scholarly endeavors.

In addition, the relationships and roles of the libraries and librarians who collaborate must inform and sustain the enrichment of these proposed digital humanities projects, as well as the relationships developed with researchers. This has significant implications for the role of the traditional liaison librarian.

The ultimate objective of this paper is to focus on the technical and relational processes of these collaborative aspects, as they might provide some valuable insights in to how we do digital humanities across borders.

How to innovate Lexicography by means of Research Infrastructures – The European example of DARIAH

Eveline Wandl-Vogt

The proposed paper focuses on the role of electronic lexicography in Europe in the framework of emerging new technologies, research infrastructures and knowledge society.

Four European research infrastructures are introduced:
DARIAH (Digital Research Infrastructure for the Arts and Humanities; www.dariah.eu), CLARIN (Common Language Resources and Technology Infrastructure; www.clarin.eu) and COST IS 1305 ENeL (European Network for Electronic Lexicography; www.elexicography.eu; http://www.cost.eu/COST_Actions/isch/Actions/IS1305).

Whereas COST ENeL is a networking project to connect people furthering innovative electronic lexicography and developing a European dictionary portal; CLARIN, DARIAH and EGI aim to develop sustainable research infrastructures both, from technological as well as social point of view.

The author is introducing DARIAHs main principles; high level scientific and technological as well as organizational and community principles.

The author exemplifies how these principles can strengthen and support to transform European Lexicography. She introduces into the main organizational units of DARIAH (e.g. working groups) maintaining lexicography.

Collaboration between these infrastructures and the strong DARIAH connection are discussed and presented on the example of lexicographical projects of the working group elexicography at the newly founded Austrian Centre for Digital Humanities (ACDH; 1.1.2015-; www.oeaw.ac.at/acdh). She focuses the most recent projects, representing the principles:
--: open access (towards open science)
--: semantic web technologies / linked data lexicography
--: standards
--: cross-sectoral collaboration (scientific community – companies – citizens – administration / politics)
--: innovation.

Concluding, she invites to participate into a global network and to further collaboration.

References

CLARIN. Common Language Resources and Technology Infrastructure; http://www.clarin.eu. Accessed 30th March 2015.

COST IS 1305 ENeL. European Network for Electronic Lexicography. http://www.elexicography.eu; http://www.cost.eu/COST_Actions/isch/Actions/IS1305. Accessed

30th March 2015.

DARIAH . Digital Research Infrastructure for the Arts and Humanities; http://www.dariah.eu. Accessed 30th March 2015.

European Commission: Research & Innovation: Infrastructures. http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=home. Accessed 30th March 2015.

Nowotny, Helga; Scott, Peter; Gibbons, Michael. 2003. `Mode 2´ Revisited. The new Production of Knowledge. Minerva 41: 179-194. http://www.uni-klu.ac.at/wiho/downloads/nowotny.pdf. Accessed 30th March 2015.

Web-based Natural Language Processing Workflows for the Research Infrastructure in Humanities

Tomasz Walkowiak and Maciej Piasecki

Text plays an important role in research in Humanities in parallel to other media. However, large amount of text used often in the research is beyond exclusively manual analysis and cause the need for the supporting automated methods for text analysis. The required processing includes frequency calculation (e.g. the number of words, parts of speech, proper names, meaning of words etc.), visualization of the extracted data, clustering as well as classification of text documents. Natural Language Processing tools can be very helpful for those tasks, if they are easily accessible and enough flexible for the researchers from Humanities. NLP software has been under development for many years and has achieved the level of practical applicability. However, several barriers blocking its widespread use in Humanities can be observed. They are even more limiting in the case of non-English languages. Tools for the processing non-English languages are very often hard to be installed.
The tools are somehow fragmented, i.e. designed for narrow specific tasks. Their integration is difficult due to the different technologies applied with complex dependencies and different data models used. In addition, processing of large texts requires huge computational power.

These problems could be overcome by making the language tools available via Web-based interfaces and providing systems for flexible but guided combination of the tools. This idea is fundamental for the CLARIN (www.clarin.eu) project and the related initiatives such as multilingual WebLicht [1] and Multiservice [2] for Polish. However, the solutions proposed so far have some limitations, like allowing only single chains (sequences) of the applications of tools. The processing offered also lacks data mining and machine learning facilities that could be very useful for many research tasks in Humanities. Therefore, we aimed at development of an open access, scalable and highly available infrastructure with various types of software interfaces that allows to build language processing research applications dedicated for Humanities. The first version presented here is focused on Polish language, but is enough flexible to be extended in the future on other languages too.

A text analysis task mostly requires running a sequence of language tools. For simple applications, like counting of number of word or Proper Name occurrences (the latter are often called Name Entities) a single sequence of tools is enough. However, for more sophisticated tasks, like text clustering, the process requires complex workflows. Therefore a special Language Processing Modelling Notation (LPMN) was developed for defining them in a way accessible to the final users. LPMN was inspired by the BPMN (http://www.bpmn.org/) and BPEL [3] graphical language used in modelling information systems. The notation allows to define the functionality of complex language tools by combining simple ones. The graphical representation of a workflow for text clustering task is presented in Fig. 1. Each input document is converted to a uniform text format. Next, each text is analysed by a part-of-speech tagger (we use WCRFT2 [4] for Polish) and then it is piped to a Name Entity recognizer (in our case it is Liner2 [5]). When the annotation phase is completed for all the texts, the feature extraction module is run (we use the tool Fextor [6]). And finally Cluto package (http://glaros.dtc.umn.edu/gkhome/views/cluto) is called to perform data clustering.

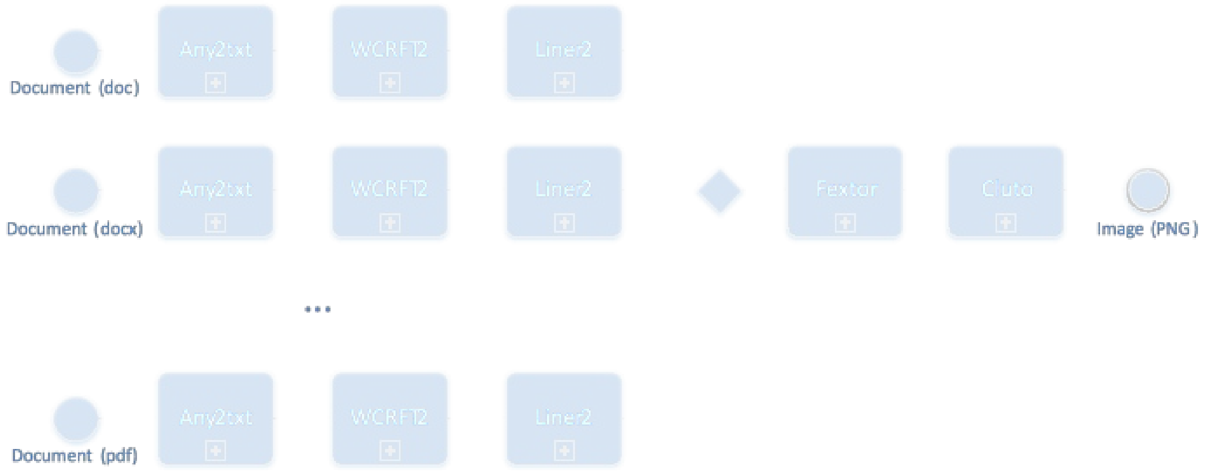


Fig 1. Text clustering workflow

The notation allows for automated execution of the workflow. That is why LPMN is defined in XML notation (like BPEL). LPMN defines also the sources of the documents to be processed (ids of files downloaded to the infrastructure, persistent identifiers from a repository or URI of a place in Internet) and destinations of results (files inside the infrastructure, our DSpace repository or e-mail address of the user).

The core of the system consists of a simple asynchronous REST service, task queues, data storage and a set of workers. The workers run language, machine learning and visualization tools. Each worker collects a task from a queue, loads data from the data storage, processes them and returns results to the data storage. The workers and the queue system allows for effective scaling of the infrastructure. Additional server grants the access from the Internet. It works as a proxy for the core system delivering a large set of different APIs. Different techniques for accessing the infrastructure including synchronous as well as asynchronous services, SOAP and REST, as well as XML and JSON are available. Such approach allows an easy integration with almost any kind of application. Moreover, the engine for running workflows described in LPMN was developed. It allows to process large corpus of text in batch like mode.

The infrastructure is monitored on different levels, starting from hardware monitoring, through virtual machines, queues and processing time of each worker.

The described infrastructure was a base for a set of application available on-line on http://ws.clarin-pl.eu and for pre-processing of texts available in CLARIN-PL digital repository, as well as for a large number of applications available on-line on http://www.clarin-pl.eu.

We plan to extend the infrastructure with support for other languages, e.g. in the case of English the work has already started this quite simple due to the availability of NLP tools. In order to improve the usability of the system for Humanities, we plan to add ontology based description of tools (modules) and an intelligent support for the definition of the workflows.

References

1. Hinrichs, M., Zastrow, T., Hinrichs, E.: WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure. . In: Proceedings of the International Conference on Language Resources

and Evaluation, pp. 489-493. European Language Resources Association (2010)
2. Ogrodniczuk, M., Lenart, M..: A multi-purpose online toolset for NLP applications. LNCS, vol. 7934, pp. 392–395. Springer-Verlag, Berlin, Heidelberg (2013)
3. OASIS. Web Service Business Process Execution Language 2.0, 2007, http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.pdf (accessed: May. 7, 2015)
4. Radziszewski, A.: A tiered CRF tagger for Polish, Intelligent Tools for Building a Scientific Information Platform. Studies in Computational Intelligence, vol. 467, pp. 215-230 (2013)
5. Marcinczuk, M.; Kocon, J., Janicki, M.: Liner2 - A Customizable Framework for Proper Names Recognition for Polish. Studies in Computational Intelligence, vol. 467, pp. 231-253 (2013)
6. Broda, B., Kędzia, P., Marcińczuk, M., Radziszewski, A., Ramocki, R., Wardyński, A.: Fextor: A feature extraction framework for natural language processing: A case study in word sense disambiguation, relation recognition and anaphora resolution. Studies in Computational Intelligence, vol. 458, pp. 41-62 (2013)

CLARIN-PL – a Polish part of the language technology infrastructure for Humanities and Social Sciences

Maciej Piasecki, Marcin Pol, Tomasz Walkowiak and Agnieszka Indyka-Piasecka

Language data, i.e. text documents, textual source materials, transcribed or recorded interviews etc. are used in different areas of Humanities and Social Sciences (H&SS). With growing amount of data available their manual analysis becomes very laborious and problematic. Language technology (i.e. language resources and language tools) could be helpful, but its application often requires specific knowledge from the area of the natural language engineering or skills in programming. H&SS researchers mostly do not posses both of them, and in fact, they do not have to. Moreover, most language tools use different data formats and have specific requirements concerning the hardware or the operating system that limits their usage by H&SS users even more. In order to improve this situation, CLARIN ERIC research consortium (www.clarin.eu) was established in 2011 by 8 European countries. The main goal of the CLARIN is the construction of a language technology infrastructure to support researchers in H&SS [1]. The infrastructure is a complex system that enables combining language tools with language resources into complex processing chains. CLARIN extends the infrastructure with research applications for processing text and speech that are built on the top of it.

CLARIN ERIC members are now 14 European countries and one intergovernmental organisation. Each member is obliged to contribute parts of the language technology infrastructure. The most typical approach to building the CLARIN-related infrastructure is a bottom-up process of linking already existing language tools and resources (LTRs). It is primarily focused on establishing accessibility and technical interoperability of LTRs. As a result, the tools and resources become accessible via Web to the users and can be combined into processing chains.
The development of CLARIN-PL (www.clarin-pl.eu) – the Polish part – has been based on a different, bi-directional model [2]. Several existing LTRs have been combined into the infrastructure, too. They provide the basic natural language processing capacity. However, H&SS user-driven requirements have been also taken into account by initiating and developing several top level research applications, i.e. partially following the top-down approach.

The top-down part is based on the idea of key users and key applications. The applications were selected on the basis of contacts established with perspective users who are researchers already active in e-Humanities and e-Social Sciences and the tendency to cover maximal variety of research areas. Due to the limited funding of CLARIN-PL (that is the case of every project) only a few applications could be chosen, but in a way making the set possibly varied across H&SS subareas and techniques required in the applications. Simultaneously, according to the bottom-up direction, LTRs that were lacking with respect to the basic Polish processing chain were identified and added to the development plan. These two approaches have been harmonised, i.e. LTR development plan has been interactively modified according to the requirements collected from the work on key applications.

During the talk, the resulting state of CLARIN-PL infrastructure will be discussed. The first year of limited operability has allowed us to draw some initial conclusions with respect to the bi-directional approach. The state of the Polish basic processing chain will be analysed in the perspective of its

usability for H&SS, as H&SS applications put some substantial requirements on it. We will identify basic language data processing schemes as they have emerged from the development of the key applications, e.g.:
a system of tools for corpus building, managing and annotating,
tools for the statistical lexical analysis of texts on the level of words and their meanings,
system supporting semi-automated semantic text corpus annotation and analysis,
tools for the comparison of corpora: static and dynamic in time,
tools for the extraction of dictionaries of terms and multiword expressions from corpora, etc.

All tools and systems have been implemented in a form of web-based applications installed in the CLARIN-PL Language Technology Centre. The architecture of the centre and its interface will be presented. Experience collected from the training sessions and tests performed with the participation of the users – H&SS researchers – will be discussed. We will also analyse the problem of the quality of services offered by language tools that goes much beyond the typical measures used during testing language tools.

In conclusion, we will try to envisage the further user needs that seem to be intensifying in time and the further LT infrastructure development for which 3-4 years construction phase is a good starting point for a fully-fledged construction.

References:

[1] Erhard Hinrichs and Steven Krauwer. The CLARIN Research Infrastructure: Resources and Tools for eHumanities Scholars. In ed. N. Calzolari et al., Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), ELRA, 2014.
URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/415_Paper.pdf
[2] Piasecki, Maciej. User-driven Language Technology Infrastructure – the Case of CLARIN-PL In Proceedings of the Ninth Language Technologies Conference, Ljubljana, Slovenia, 2014.
URL: http://nl.ijs.si/isjt14/proceedings/isjt2014_01.pdf

JADH2015 Conference Booklet
Published by the Center for Informatics in East Asian Studies,
Institute for Research in Humanities, Kyoto University,
47 Higashiogura-cho, Kitashirakawa Kyoto 606-8265, JAPAN
Responsible Editor: Christian Wittern
ISBN 978-4-9906708-7-0