

# JADH 2021

“Digital Humanities and COVID-19”

September 6-8, 2021

The University of Tokyo, JAPAN



<https://www.hi.u-tokyo.ac.jp/JADH/2021/>

## The 11th Conference of Japanese Association for Digital Humanities

**Proceedings of JADH conference, vol. 2021**

Organized by

Organizing Committee, Japanese Association for Digital Humanities

Hosted by

Historiographical Institute, The University of Tokyo

Co-organized by

International Institute for Digital Humanities

Supported by

IPSJ SIG Computers and the Humanities

Japan Art Documentation Society

Japan Association for English Corpus Studies

Japan Society for Digital Archives

Japan Society for Information and Media Studies

Japan Society of Information and Knowledge

The Mathematical Linguistic Society of Japan

Proceedings of JADH conference, vol. 2021

Edited by Historiographical Institute, The University of Tokyo

Copyright © 2021 by the Japanese Association for Digital Humanities

Published by Historiographical Institute, The University of Tokyo

3-1, Hongo 7-chome, Bunkyo-ku, Tokyo 113-0033, JAPAN

<https://www.hi.u-tokyo.ac.jp/>

Online edition: ISSN 2432-3144

Print edition: ISSN 2432-3187

## Table of Contents

<b>Table of Contents.....</b>	<b>3</b>
<b>JADH 2021 Committees .....</b>	<b>9</b>
<b>Time Table.....</b>	<b>11</b>
<b>[Plenary – 1] .....</b>	<b>12</b>
<b>Keener than Connoisseurs’ Eyes: Analysis and Experience of Ancient Art through Virtual Reality (VR) .....</b>	<b>12</b>
<i>Kyoko Haga</i>	
<b>[Plenary – 2] .....</b>	<b>14</b>
<b>(Dis)Connections in Digital Japanese Studies .....</b>	<b>14</b>
<i>Paula R. Curtis</i>	
<b>Style Comparative study of Japanese medieval picture scrolls focusing on landscapes using GM Method with IIF Curation Platform .....</b>	<b>16</b>
<i>Chikahiko Suzuki, Akira Takagishi, Asanobu Kitamoto</i>	
<b>Book Barcoding for Differential Reading -Application to Woodblock-printed Books in the Bukan Complete Collection- .....</b>	<b>22</b>
<i>Asanobu Kitamoto</i>	
<b>Digital technologies and the spatial organisation of exhibitions: Interactive art as reflective experience .....</b>	<b>28</b>
<i>Marianna Charitonidou</i>	
<b><i>The Re-centered Text: Digital restructuring in Amira Hanafi’s A Dictionary of the Revolution.....</i></b>	<b>31</b>
<i>David Thomas Henry Wright</i>	
<b>Experimental LDA Topic Modelling of Tennyson’s Epic Poems.....</b>	<b>34</b>
<i>Iku FUJITA</i>	
<b>A study on the readerly aspects of Electronic Poetry through Cognitive Poetics .....</b>	<b>45</b>
<i>Mariyam Nancy J, David Arputharaj</i>	

<b>Architectural Drawings Exposed and the Effect of Digitization: The Rise of Artefactual Value vs the Democratization of Knowledge.....</b>	<b>47</b>
<i>Marianna Charitonidou</i>	
<b>Intersectionality and Digital Humanities in the Teaching of Architectural History: Diversity in the Dissemination of Knowledge .....</b>	<b>48</b>
<i>Marianna Charitonidou</i>	
<b>Multilingual word embeddings and low resources: identifying influence in Antiquity .....</b>	<b>51</b>
<i>Marianne Reboul, École Normale Supérieure de Lyon</i>	
<b>Skin Deep: Exploring Ideals of Japanese Beauty through Social Media .....</b>	<b>55</b>
<i>Amy Grace Metcalfe, Emily Ohman</i>	
<b>Analyzing “Mechanisms” in the British National Corpus .....</b>	<b>59</b>
<i>Yuki Sugawara</i>	
<b>One Challenge, Not Two Problems: Regular Expressions for Researching a Single-Author Corpus.....</b>	<b>62</b>
<i>Dr. Robert W. Williams</i>	
<b>Picking out Arabian Names from <i>Fahrasa</i> by Ja‘far b. Idrīs al-Kattānī without Reading Arabic.....</b>	<b>66</b>
<i>Yuri Ishida, Kensuke Baba</i>	
<b>POS tagging for Vedic Sanskrit using deep learning.....</b>	<b>69</b>
<i>Yuzuki Tsukagoshi</i>	
<b>Spectral analysis for identifying octave playing in piano works .....</b>	<b>72</b>
<i>Mai Takahashi, Michikazu Kobayashi, Ikki Ohmukai</i>	
<b>Token-based semantic vector space model for classic poetic Japanese.....</b>	<b>77</b>
<i>Xudong Chen, Hilofumi Yamamoto, and Bor Hodošček</i>	
<b>Open source datasets of the Hachidaishū for the research of classical Japanese poetic vocabulary.....</b>	<b>82</b>
<i>Hilofumi Yamamoto, Bor Hodošček</i>	

<b>Exploring Metadata Quality Issues in Non-English Corpora: Preliminary Assessments of HathiTrust Records of Late Imperial Chinese Books .....</b>	<b>88</b>
<i>Wenyi Shang, Jacob Jett, J. Stephen Downie</i>	
<b>Dataset Construction for Cross-genre Plot Structure Extraction.....</b>	<b>93</b>
<i>Hajime Murai, Shuuhei Toyosawa, Takayuki Shiratori, Takumi Yoshida, Shougo Nakamura, Yuuri Saito, Kazuki Ishikawa, Sakura Nemoto, Junya Iwasaki, Akiko Uda, Shoki Ohta, Arisa Ohba, Takaki Fukumoto</i>	
<b>Basic Plot Structure in the Adventure and Battle Genres .....</b>	<b>97</b>
<i>Yuuri Saito, Takumi Yoshida, Shougo Nakamura, Kazuki Ishikawa, Shoki Ohta, Arisa Ohba, Takaki Fukumoto, Hajime Murai</i>	
<b>Construction of ShiJi Spatiotemporal Information Platform on the Framework of Research-oriented Knowledge Bases .....</b>	<b>101</b>
<i>Jung-Yi Tsai, Pi-Ling Pai, Hsiung-Ming Liao, You-Jun Chen, Richard Tzong-Han Tsai*, I-Chun Fan</i>	
<b>Cross-genre Plot Analysis of Detective and Horror Genres .....</b>	<b>106</b>
<i>Junya Iwasaki, Shuuhei Toyosawa, Kazuki Ishikawa, Shoki Ohta, Hajime Murai</i>	
<b>Using Moodle as a Multi-Modal Tool for Ainu Language Education .....</b>	<b>111</b>
<i>Matthew Cotter, Takayuki Okazaki, Jennifer Teeter</i>	
<b>An Attempt at Creating Integrated Retrieval for Chinese Excavated Materials: An Implementation of a Search Function across Interpretations of Ancient Characters .....</b>	<b>114</b>
<i>Shumpei Katakura</i>	
<b>Collecting Canons: Comparing Guodian and Mawangdui Laozi Texts with the Dead Sea Scrolls.....</b>	<b>117</b>
<i>Janelle Peters</i>	
<b>Development of Database for Japanese Conversation Patterns: an observation from noun phrases ending with focus particle "mo (also)" .....</b>	<b>120</b>
<i>Mika Ebara, Hilofumi Yamamoto</i>	

<b>Drug-focused text summarization of coronavirus-related articles for the discovery of COVID-19 therapies .....</b>	<b>124</b>
<i>Setsuro Matsuda</i>	
<b>Reconstruction and Utilization of Text Data Using TEI: Case study of the Shibusawa Eiichi Denk Shiryō .....</b>	<b>126</b>
<i>Boyoung Kim, Satoru Nakamura, Yuta Hashimoto, Naoki Kokaze, Sayaka Inoue, Toru Shigehara, Kiyonori Nagasaki</i>	
<b>Development of a support system for extracting mentioned bibliographical data from the <i>Encyclopédie</i> entries .....</b>	<b>130</b>
<i>Satoru Nakamura, Ayano Kokaze, Yoshiho Iida, Naoki Kokaze, Tatsuo Hemmi</i>	
<b>Platformed reflections on the Pandemic: Covid-19 and Electronic Literature.....</b>	<b>134</b>
<i>Anna Nacher, Søren Bro Pold, Scott Rettberg</i>	
<b>Digital Humanities and the way forward for ethnographic research: What we learned from Covid-19? .....</b>	<b>139</b>
<i>Deepika Kashyap</i>	
<b>Virtual Communities and Post-Pandemic Possibilities: Animal Crossing New Digital Humanities .....</b>	<b>141</b>
<i>Quinn Dombrowski, Elizabeth Grumbach, Merve Tekgürler</i>	
<b>Building Web Corpus of Old Nubian with Interlinear Glossing as Digital Cultural Heritage for Modern-Day Nubians.....</b>	<b>144</b>
<i>So Miyagawa, Vincent W.J. van Gerven Oei</i>	
<b>Development of data-driven historical information research infrastructure at the Historiographical Institute in the University of Tokyo.....</b>	<b>148</b>
<i>Satoru Nakamura, Taizo Yamada</i>	
<b>Compilation of Semantic Data Archive: A New Method of Learning “Local Culture” .....</b>	<b>152</b>
<i>Kwangwoo Kim, Soohyeon Kim</i>	
<b>Towards a Structured Description of the Contents of the Taisho Tripitaka.....</b>	<b>161</b>

*Yoichiro Watanabe, Kiyonori Nagasaki, Hyunjin Park, Yifán Wáng, Tomohiro Murase, Masayoshi Watanabe, Norimichi Yajima, Yoshihiro Sato, Yūi Sakuma, Xinxing Yu, Masahiro Shimoda, Ikki Ohmukai*

**Classification of face images in the frontispiece paintings of Sutra copies in gold ink on indigo paper by deep convolutional neural networks ..... 164**

*Toshiaki Aida, Tomomi Kobayashi, Aiko Aida*

**The difference in transitional process between Western instrumental and vocal music ..... 169**

*Daisuke Miki, Akihiro Kawase, Kenji Hatano*

**e-Sukhāvati: An Innovative Digital Platform for Studying the *Smaller Sukhāvavīyūha* ..... 172**

*SIU Sai-yau*

**New Possibilities of Digital Publishing and Online Exhibition— A Case Study of the Website “Reflections on COVID-19” ..... 178**

*Lin, Wen Jiun*

**Sonifying the pandemic – innovative approaches towards data interaction and engagement formats for scientific, educational and artistic purposes ..... 192**

*Michael Stark, Amelie Dorn, Renato Rocha Souza*

**Thailand Towards Digitization– the past, the present, the future and gray digital gap ..... 196**

*Saiyud Moolphate, Nadila Mulati, Thin Nyein Nyein Aung, Motoyuki Yuasa, Myo Nyein Aung*

**Wikidata as a Low-tech Solution to Leverage Semantic Technologies and A Case Study of CBDB ID’s Reconciliation with Wikidata ..... 199**

*Fudie Zhao*

**[Workshop – 1] ..... 203**

歴史学におけるデータ共有, 統合化, 多角的協働 ..... 203

**[Workshop – 2] ..... 205**

海外 DH 教育動向調査 ..... 205



## JADH 2021 Committees

### Program Committee:

- Paul Arthur (Edith Cowan University, Australia)
- Marcus Bingenheimer (Temple University, USA)
- James Cummings (Newcastle University, UK)
- J. Stephen Downie (University of Illinois, USA)
- Øyvind Eide (University of Cologne, Germany)
- Makoto Goto (National Museum of Japanese History, Japan)
- Shoichiro Hara (Kyoto University, Japan)
- Yuta Hashimoto (National Museum of Japanese History, Japan), Chair
- Bor Hodošček (Osaka University, Japan)
- JenJou Hung (Dharma Drum Institute of Liberal Arts, Taiwan)
- Jieh Hsiang (National Taiwan University, Taiwan)
- Akihiro Kawase (Doshisha University, Japan)
- Nobuhiko Kikuchi (Kansai University, Japan)
- Asanobu Kitamoto (ROIS-DS Center for Open Data in the Humanities / National Institute of Informatics, Japan)
- Maciej Eder (Pedagogical University of Kraków, Poland)
- Yoko Mabuchi (National Institute for Japanese Language and Linguistics, Japan)
- Charles Muller (University of Tokyo, Japan)
- Hajime Murai (Future University Hakodate, Japan)
- Kiyonori Nagasaki (International Institute for Digital Humanities, Japan)
- Satoru Nakamura (University of Tokyo, Japan)
- Chifumi Nishioka (Kyoto University, Japan)
- Ikki Ohmukai (University of Tokyo, Japan)
- Geoffrey Rockwell (University of Alberta, Canada)
- Martina Scholger (University of Graz, Austria)
- Masahiro Shimoda (University of Tokyo, Japan)
- Raymond Siemens (University of Victoria, Canada)
- Tomoji Tabata (Osaka University, Japan)
- Ruck Thawonmas (Ritsumeikan University, Japan)
- Toru Tomabechi (International Institute for Digital Humanities, Japan)
- Kathryn Tomasek (Wheaton College, USA)
- Ayaka Uesaka (Osaka University, Japan)
- Raffaele Vighianti (University of Maryland, USA)
- Christian Wittern (Kyoto University, Japan)
- Taizo Yamada (University of Tokyo, Japan)

- Natsuko Yoshiga (Saga University, Japan)

**Local Organizers:**

- Hiroshi Hakoishi (Historiographical Institute, The University of Tokyo)
- Kanako Hirasawa (Historiographical Institute, The University of Tokyo)
- Satoshi Inoue (Historiographical Institute, The University of Tokyo)
- Kiyonori Nagasaki (International Institute for Digital Humanities)
- Satoru Nakamura (Historiographical Institute, The University of Tokyo)
- Ikki Ohmukai (Graduate School of Humanities and Sociology, The University of Tokyo)
- Ayako Shibutani (Historiographical Institute, The University of Tokyo)
- Toru Tomabechi (International Institute for Digital Humanities)
- Taizo Yamada (Historiographical Institute, The University of Tokyo) - Chair
- Hidenori Watanabe (Interfaculty Initiative in Information Studies, The University of Tokyo)

## Time Table

### September 6 (Mon), Day 1

10:30-14:00	Workshop 1
15:00-17:00	Workshop 2

### September 7 (Tue), Day 2

10:00-10:30	Opening
10:30-12:00	Long Paper 1: Visuality (Room A)
10:30-12:00	Long Paper 2: Literature and Poetry (Room B)
12:00-13:00	Board Meeting
13:00-14:00	Plenary Talk I: Kyoko Haga (Room A)
14:30-16:00	Long Paper 3: Architecture (Room A)
14:30-16:00	Short Paper 1: Text Analysis (Room B)
16:30-18:00	Poster
18:00-19:30	Free Talk

### September 8 (Wed), Day 3

9:30-10:30	Plenary Talk II (Room A)
11:00-12:00	Long Paper 4: Historical Texts (Room A)
12:00-13:00	AGM (JADH Annual General Meeting, Room A)
13:30-15:00	Long Paper 5: DH and the Pandemic (Room A)
13:30-15:00	Short Paper 2: Archiving and Analysis of Cultural Heritages (Room B)
15:30-17:00	Short Paper 3: Digitization under the Pandemic (Room A)
17:00-17:30	Closing (Room A)

## [Plenary – 1]

# Keener than Connoisseurs' Eyes: Analysis and Experience of Ancient Art through Virtual Reality (VR)

Kyoko Haga<sup>1</sup>

### Bio

Kyoko Sengoku-Haga is Professor of Art History with the Center for Evolving Humanities in the Graduate School of Humanities and Sociology at the University of Tokyo. Her research focuses on both religious and technical aspects of Western Classical art, and regarding the latter, from 2007 she has been conducting a joint research project with Prof. Katsushi Ikeuchi and Prof. Takeshi Oishi in the Computer Vision Laboratory of the Institute of Industrial Science at the same university. She is also a project member of the Virtual Reality Educational Research Center and a fellow of the Art Center, both of which are organizations within the University of Tokyo. She has written books and articles on Greek and Roman Art, including *The Ancient Sculpture of Rhodes* (2006, Collegium Mediterranistarum Herend Award in 2007) and *History of Western Art*, vol. 1 (2017, main author), and supervised several special exhibitions on ancient Greek and Roman art such as *A Journey to the Land of Immortals: Treasures of Ancient Greece* (Tokyo National Museum, 2016) and *Roman Wall Paintings of Pompeii* (Mori Arts Center Gallery, Tokyo, 2016). From 2017 she is promoting a VR Art Appreciation project of special art exhibitions in collaboration with Asahi Shimbun-sha.

### Abstract

In the field of art, VR technique (in its broadest sense) is applied in various ways, such as to record and reproduce works of art for various purposes and to analyze them for research purposes, as well as to create new works.

The importance of recording and reproducing works of art has long been recognized. At first, this was in terms of the cultural heritage, which is in danger of damage or destruction. Now, many projects of recording works of art as high-resolution images are going on in the world and many museums, sometimes in collaboration with Google Arts and Cultures, have opened their collections of data on the internet. In addition, during these years under the COVID-19 crisis, some special exhibitions have been recorded as 3D Walkthroughs although in a rather simple way. If we can develop a method of archiving special exhibitions as VR content that is realistic enough to be appreciated as

---

<sup>1</sup> Graduate School of Humanities and Sociology, the University of Tokyo

art, it may help those people who have difficulty accessing real museums including high school students of remote rural areas and old people with physical challenges. It may solve the regional disparities in art education and enrich people's lives in an aging society.

Regarding the use of VR, or 3D data, for the purpose of art studies, it is extremely useful for analyzing ancient art, because the conserved artworks are not in abundance, and in most cases, the condition is quite poor. With the help of 3D models, we can reconstruct fragmentary objects correctly at least in respect to its geometric aspect. However, that is not all; the digital eye can analyze art objects better than any skilled connoisseurs' eyes.

Taking an example of our research on ancient Greek and Roman art, 3D shape comparison technique enables us to examine the precision of making copies of famous works. To explain the method briefly, firstly, ancient statues are scanned with 3D laser-range sensors with an accuracy of  $\pm 50\mu\text{m}$  to create very precise 3D models. Next, to compare their shapes, two 3D models are aligned together using the ICP algorithm. Lastly, to visualize the gap of the two 3D models, the distances between them are colored. We compared four Roman copies of a single Greek original masterpiece, the Doryphoros (the Spear-Bearer), created by Polykleitos in the 5<sup>th</sup> century BCE. As a result, we proved that Roman sculptors were capable of copying Greek originals quite precisely both in bronze and in marble. Even in marble copies concerning the heads and feet of statues, we observed that the errors of copies are within 2mm.

With the same method, we have revealed the actuality of a Greek sculptor's creation process as well. Comparing the Doryphoros of Polykleitos with another of his works, the Diadoumenos (the youth tying a fillet around his head), we found that the great sculptor reused the face and foot models of the former work for the latter. He must have kept the master model of the Doryphoros in his workshop, presumably to repeat its "Canon" ("standard", probably a combination of body part proportions) in a new work. It suggests the possibility of distinguishing his unidentified works from his contemporaries' works. Actually, in the case of the famous statuary group of three Wounded Amazons, one of the three was to be attributed to Polykleitos but without any reliable evidence scholars continued the discussion for a hundred years. However, we have succeeded to get to the answer by the clear results of the 3D shape comparison.

## [Plenary – 2]

### (Dis)Connections in Digital Japanese Studies

Paula R. Curtis<sup>1</sup>

#### Bio

Paula R. Curtis is a historian of medieval Japan. She is presently Postdoctoral Fellow and Lecturer in History with the Terasaki Center for Japanese Studies at University of California, Los Angeles. Her current book project, *The Casters of Kawachi: Artisans and the Production of Medieval Japan*, focuses on metal caster organizations and their relationships with elite institutions from the twelfth to sixteenth centuries. She also works on the history of documentary forgery in premodern Japan and recently published “An Entrepreneurial Aristocrat: Matsugi Hisanao and the Forging of Imperial Service in Late Medieval Japan” with *Monumenta Nipponica*.

In addition to historical work on premodern Japan, Dr. Curtis collaborates in, leads, and produces numerous online projects. Most recently, she is working to develop *Japan Past & Present*, a new digital platform project sponsored by the Yanai Initiative for Globalizing Japanese Humanities at the University of California, Los Angeles and Waseda University. This project is a joint effort between UCLA and Waseda to promote the study of Japan globally, across disciplinary and geographic borders. Dr. Curtis is a collaborator with Digital Humanities Japan, an international and interdisciplinary community of scholars and professionals interested in utilizing digital methods, tools, and resources for Japanese Studies, and hosts the blog *What can I do with a B.A. in Japanese Studies?* In addition, she curates several online resources, such as job advertisement data and visualizations in East Asia Studies and a database of Digital Resources and Projects on East Asia. More information can be found on her website: <http://prcurtis.com/>

#### Abstract

As academics, we are primed to question ourselves along with our research subjects. Many who attend this conference have likely had a moment where they have pondered “Am I a digital humanist?” A question that is no less daunting than the perpetual “What is digital humanities?” Why does claiming DH scare many of us, and why does it matter to still take the leap? This talk will address the stumbling blocks that many early career researchers

---

<sup>1</sup> Postdoctoral Fellow and Lecturer in History, Terasaki Center for Japanese Studies, University of California, Los Angeles

(and, indeed, many others) take in discovering the possibilities and limitations of digital methods. Speaking from personal experience in my own recent, digital beginnings exploring documentary culture and socioeconomic networks among artisans and their patrons in medieval Japan, I will highlight not only the connections we make in our research, but also the (dis)connections within our scholarly circles that have led to the comparatively slow growth of digital Japanese Studies in North America and elsewhere. I will offer an overview of the state of the field in North America, including conferences, workshops, publications, future projects, and, more importantly, resources for community building and collaboration that remain underutilized. A truly global DH Japan community has yet to blossom, however. One core reason is that, though our field encounters unique obstacles because of the Eurocentric and monolingual digital environments that dominate DH, the real challenges facing our progress in digital Japanese Studies are not technological, but social and institutional. This talk will both survey these issues and offer suggestions on how we might productively move forward.

# Style Comparative study of Japanese medieval picture scrolls focusing on landscapes using GM Method with IIF Curation Platform

Chikahiko Suzuki<sup>1</sup>, Akira Takagishi<sup>2</sup>, Asanobu Kitamoto<sup>1</sup>

## Introduction

Illustrations of picture scrolls in medieval Japan are divided into two elements. One is the human figures, who appear in the story. The other is a landscape, which constitutes the stage where people are drawn. Both elements are important for analyzing the style of painters.

We have already performed an analysis about facial expressions of human figures using GM method [1][2] on Platform (ICP) [3][4]. By analyzing the details and the whole, we were able to reinforce previous research and gain new findings. In addition, GM method was able to compare a large number of facial expressions that were difficult with the physical method such as using cards or photos. It has also become possible to share research results in a reusable format.

This time we focus on landscapes, especially depiction of plants and waters. Same as facial expressions, analyzing landscapes in picture scrolls can clarify the situation of production. We also aim to show the potential of GM method that can be applied not only to facial expressions but also to various elements.

## Materials and Methods

The target material is *Yugyo Shounin Engi-Emaki Shojo-Kouji Kouhon* 遊行上人縁起絵巻 清浄光寺 甲本 (*Kouhon*) archived in *Shojo-Kouji Temple*. This is a 10-volume picture scroll depicting the foundation of the *Jishu* sect of Buddhism in the *Kamakura* period (12th to 14th). *Kouhon* is one of the manuscripts of *Yugyo Shonin Engi Emaki* whose original version has been lost.

The production of *Kouhon* has been studied by many scholars. Iwahashi proposed a hypothesis that *Kouhon* is a mixture of three styles (ABC) [5]. Our study on the analysis of facial expressions with the GM method also supports this hypothesis.

Combining our knowledge about the production of picture scrolls in medieval Japan, ABC styles likely correspond to three workshops, not three individual painters. Each workshop consisted of the master painter who is in charge of the main part and several assistant painters who draw the secondary part such as the landscape.

---

<sup>1</sup> ROIS-DS Center for Open Data in the Humanities / National Institute of Informatics

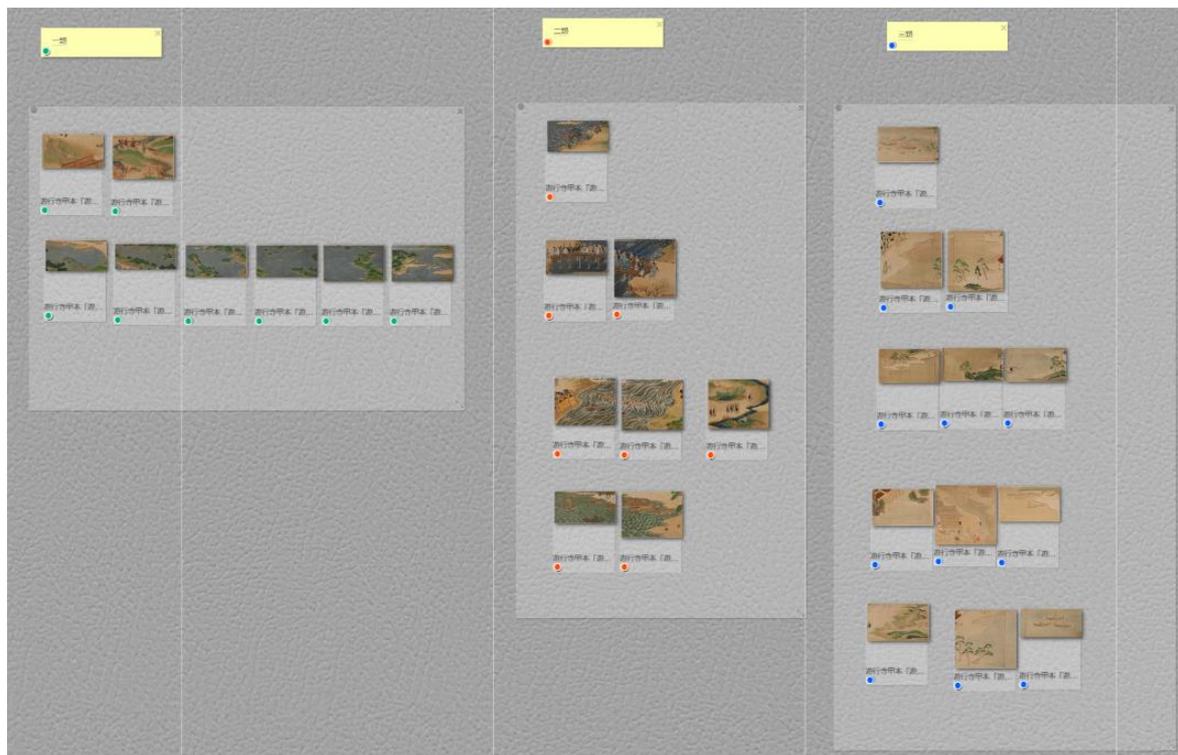
<sup>2</sup> The University of Tokyo

To compare different styles, we applied the GM method for organizing and analyzing digital images from the *Kouhon*. The GM method consists of two steps. The first step is “curation” in which we crop a part of images and add metadata. The second step is “arrangement” of images using the metadata. The GM method could be performed in an analog way using scissors, glue or photo cards, the digital version of the GM method using the IIF Curation Platform enables us to perform a large-scale analysis for any IIF materials with support for sharing the results in a reusable format.

For analysis, we cropped 180 plants and 28 depicted waters from *Kouhon*. We then added metadata such as "theme" and "source (卷 volumes / 段 scenes)". We used the IIF Curation Board [6] to arrange landscape elements by ABC styles identified in the previous research to compare different styles (Fig1, 2).



**Figure 1: Plants arranged with IIF Curation Board**  
(Seal color is green = painter A, red = painter B, blue = painter C)



**Figure 2: Water arranged with IIF Curation Board**

**(Seal color is green = painter A, red = painter B, blue = painter C)**

## Results

First, we show the results of analysis on the water. Although water has only 28 elements in total, the three styles are clearly different. Water classified as A has thin lines in both the sea and river, and the waves are regularly drawn. The water drawn by B is decorative, with waves shaped like scales on the sea and complex lines representing the flow on the river. It seems that the sea and the river have different styles. However, in every scene, the tip of the wave is drawn in a way called *Warabite-jou* 蕨手状, so we can read a common painting style. Although C is drawn using some shades, it is less drawn than the other two styles.

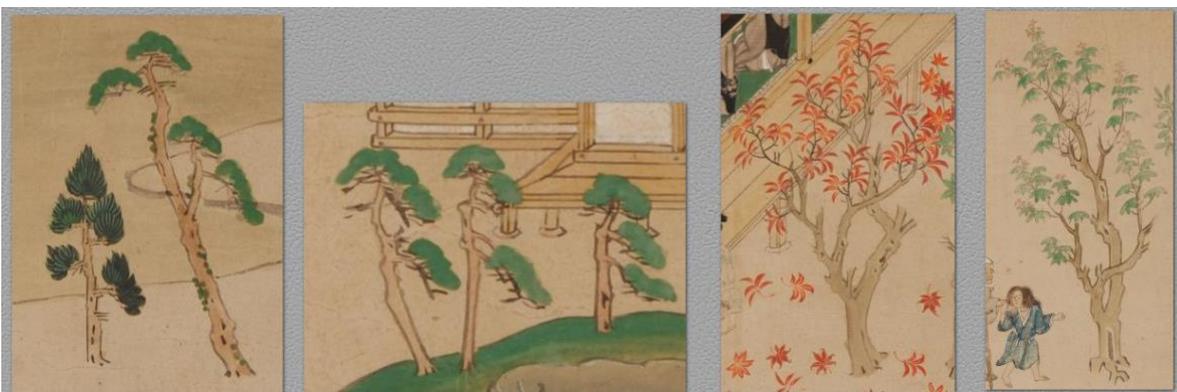
Second, we show the result of analysis on the plants. Japanese picture scrolls use “pine” to represent an evergreen tree, and “maple” as a deciduous tree. From the way of drawing these typical trees, the three styles are clearly different. A draws a tree with well-proportioned style with a straight centerline. Another feature is that the dotted moss *Tentai* 点苔 that attaches to trees is regularly arranged. The impression is classic in the Middle Ages (Fig3). On the other hand, B draws a tree with a thick outline and various forms with many curves. It is also characterized by moss with irregular size and arrangement. B draws bright and huge flowers that do not appear in other painters' parts (Fig. 4). C has a pale style, with less overlapping branches and a flat tree shape (Fig. 5).



**Figure 3: Classic style trees by A**

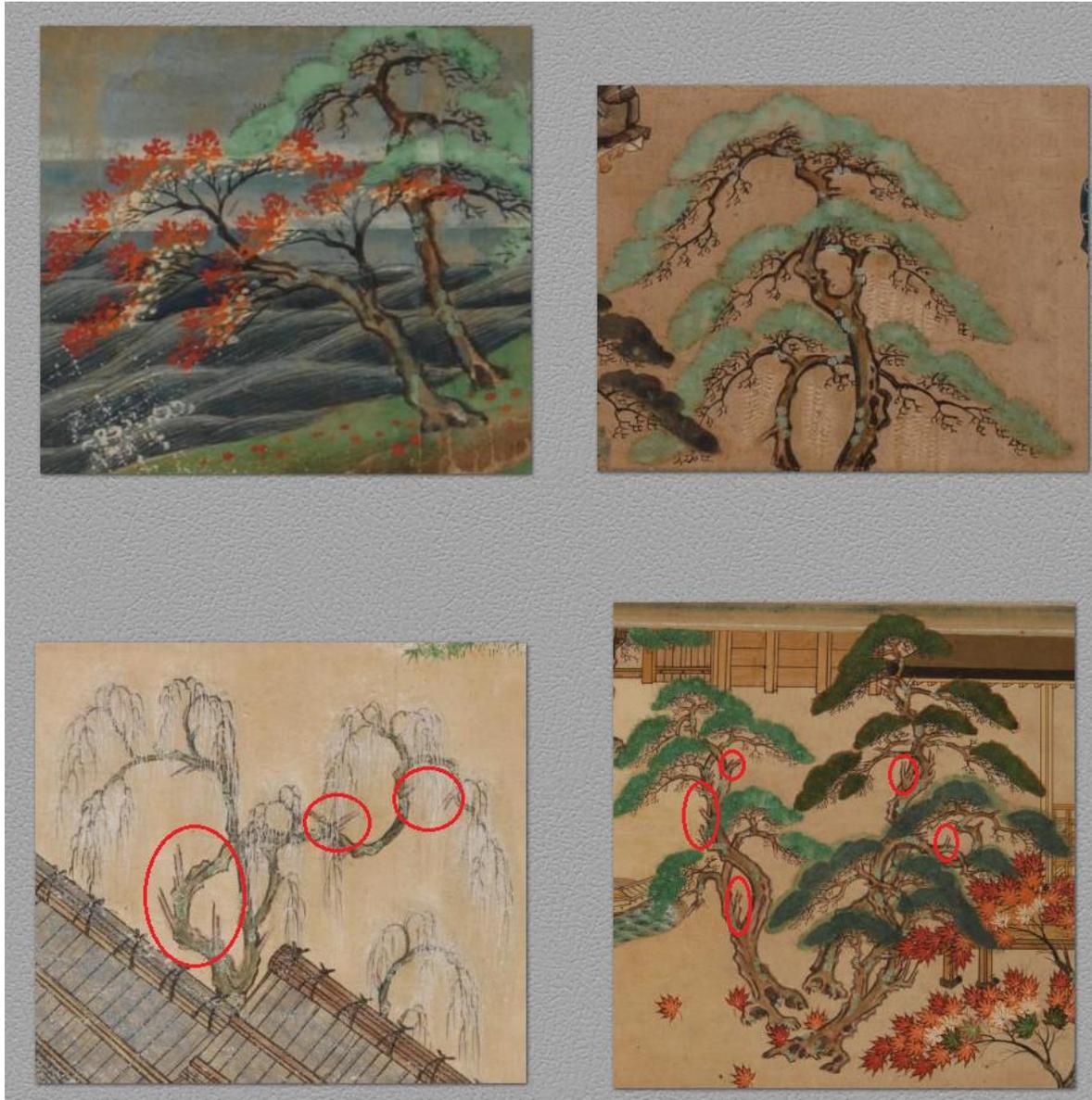


**Figure 4: Various forms of trees and flowers by B**



**Figure 5: Pale and flat style by C**

In the next step, we focused on the details of the plant. We can observe different manners within ABC styles. B has common features, but the way of drawing branches differs depending on the volume and scene (Fig6). Due to the difference in manner, we can find the hands of multiple painters. A has a small number of plants, C is a pale and flat depiction, so it is difficult to find a different manner in each. Even so, it is possible to assume the hands of multiple painters from the difference in the amount of drawing.



**Figure 6: Variations of the trees by B.**

**Characteristic broken branches that do not appear in the first half (upper), appear in the second half (lower). Emphasized with a red circle.**

### **Discussion**

By analyzing the landscape of *Kouhon* using the GM method, we support previous research that identified three styles. Furthermore, we found different manners in each style. This finding leads to the restoration of the production situation. This also confirms the hypothesis that master and assistant painters performed production in a team.

We also showed that the GM method is effective not only for human figures but also for landscapes. This method may be extended to other elements that make up the landscape such as soil, buildings, and small tools. We are planning to apply this method to the comparison of multiple works to help clarify the lineage of manuscripts.

**Reference**

- [1]. Chikahiko Suzuki, Akira Takagishi, Alexis Mermet, Asanobu Kitamoto, Jun Homma. Analysis of difference between male and female facial expressions in Japanese picture scrolls using GM Method with IIF Curation Platform, JADH2020, pp.90-95, 2020
- [2]. Chikahiko Suzuki, Akira Takagishi, Jun Homma, Alexis Mermet, Asanobu Kitamoto. Difference between male and female in medieval Japanese picture scrolls -Style comparative study for Yugyo Shounin Engi-Emaki using GM Method with IIF Curation Platform-, Jinmoncom Symposium 2020, pp.67-74, 2020
- [3]. IIF Curation Platform <http://codh.rois.ac.jp/icp/index.html.en> Accessed on 2021-08-20.
- [4]. Asanobu Kitamoto, Jun Homma, Tarek Saier. IIF Curation Platform: Next Generation IIF Open Platform Supporting User-Driven Image Sharing, Jinmoncom Symposium 2018, pp.327-334, 2018
- [5]. Haruki Iwahashi. Yugyo-Shonin-Engi-Emaki; Seijoko-Ji-Bon ni tsuite. *Ars buddhica* (185), pp.51-59, 1989
- [6]. IIF Curation Board <http://codh.rois.ac.jp/software/iif-curation-board/> Accessed on 2021-08-20.

# **Book Barcoding for Differential Reading -Application to Woodblock-printed Books in the Bukan Complete Collection-**

**Asanobu Kitamoto<sup>12</sup>**

## **1. Introduction**

This paper proposes a method called “book barcoding” for the bibliographic study of woodblock-printed books. The goal is to realize “differential reading” [4,5], which is a method for highlighting visual differences between different books with the help of computational algorithms. Textual comparison of different books has been extensively studied, but the visual comparison is still dependent on the manual side-by-side comparison, which is both time-consuming and highly unreliable. To solve this problem, we propose “book barcoding,” a method that combines several computational algorithms with the following research contributions. First, we have extended a page-by-page collation method, already proposed by the authors, to a book-by-book collation method with the help of the Gale-Shapley algorithm [3]. Second, we developed a tool called “vdiff.js” for highlighting differences in the page-by-page collation. Third, we constructed a platform for differential reading of the Bukan Complete Collection on the book-by-book collation.

## **2. Book Barcoding Method**

### **What is book barcoding?**

Book barcoding is inspired by “DNA barcoding” [7] in the field of biodiversity research for specimen identification and species discovery. DNA barcodes are unique DNA sequences, placed in the Barcode of Life Data Systems (BOLD) database - an online workbench that includes a reference library of DNA barcodes to assign identities to sequences of unknown origin. This idea motivated us to develop a similar framework for woodblock-printed books to develop a reference library of features of known page images and identify the book of unknown origin by finding the best match in the library. The book barcoding consists of collations in two levels, namely the page-by-page collation, and the book-by-book collation.

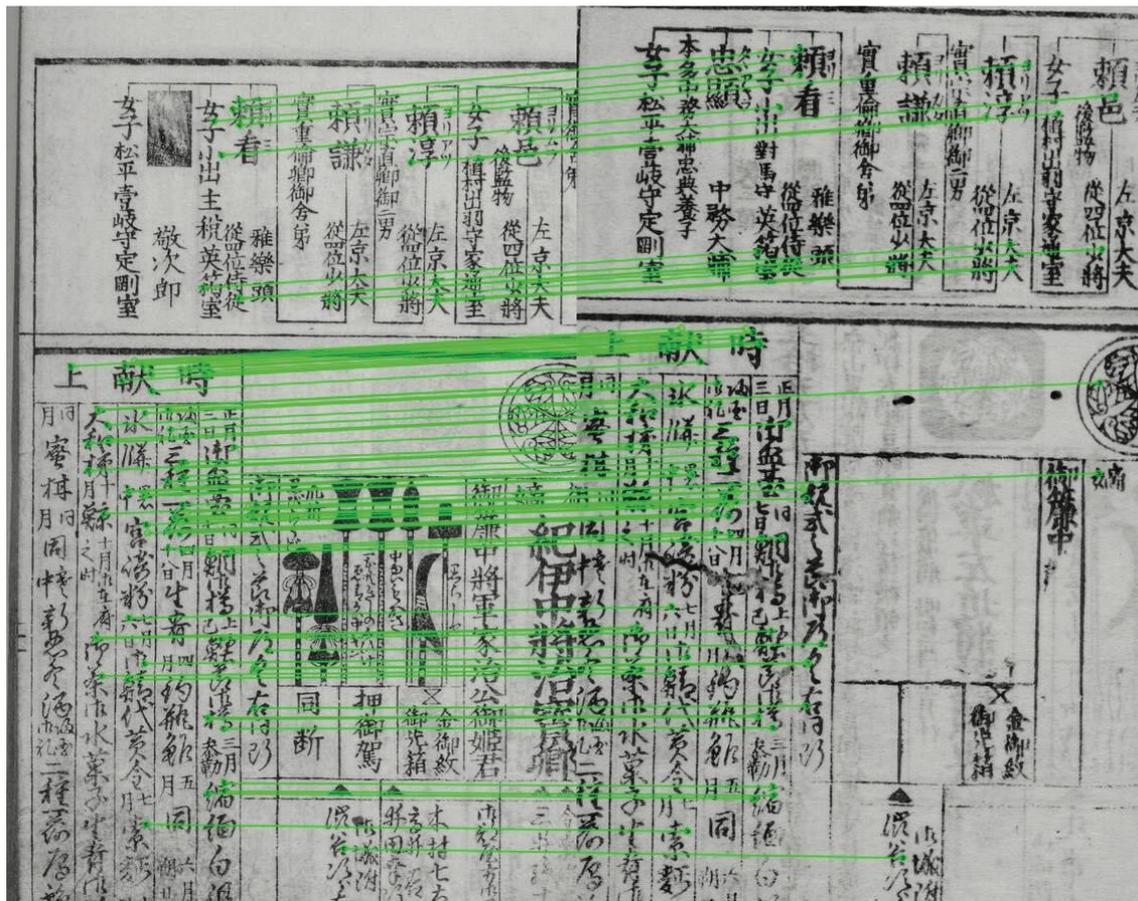
### **Page-by-page collation**

---

<sup>1</sup> ROIS-DS Center for Open Data in the Humanities (CODH)

<sup>2</sup> National Institute of Informatics

We have already proposed a method for the page-by-page collation [6]. First, we apply the AKAZE feature detectors in OpenCV<sup>3</sup> and extract keypoints and descriptors from a page image to store them in the database. Second, given the pair of page images, we fetch keypoints and descriptors from the database, match them using Hamming distance as the distance metric between descriptors, and estimate a projective transformation matrix using the RANSAC algorithm [2]. Third, we count the number of inlier keypoints to estimate the goodness of matching between two images. Green lines in Figure 1 shows corresponding inlier keypoints.



**Figure 1: Matching of keypoints for the page-by-page collation. The green lines show corresponding inlier keypoints between two images.**

### Book-by-book collation

This step focuses on selecting the best matching of pages between two books. This problem can be formulated as a stable marriage problem, which was originally proposed for finding the best match between a group of men and a group of women, assuming that each member has an ordering of preference for a member in another group. We define the preference as the number of inlier keypoints, where a higher number means a higher preference. We use the Gale-Shapley algorithm, a classical algorithm for computing the

<sup>3</sup> <https://opencv.org/>

best matching, to obtain the list of page pairs between two books, based on the “barcode” of each page.

### 3. Book Barcoding Method

#### Processing Workflow

To prepare for the differential reading, we established a workflow as follows. First, we selected 336 Bukan books in the “Bukan Complete Collection” released from ROIS-DS Open Data Center for Humanities (CODH), derived from digitized images from the National Institute of Japanese Literature (NIJL). The form of books is either portrait or landscape. A portrait book was captured as an image of facing pages, so we split it into two images so that one image captures the paper region of a half sheet. On the other hand, a landscape book was captured as a half sheet, so we only cropped the paper region. The dataset has 143,616 images, including 111,114 portrait images (from 55,557 original images) and 32,502 landscape images. We then applied the AKAZE feature detector to obtain 67,071,993 keypoints, or on average, about 467 keypoints for each image. Next, we applied the page-by-page collation using the RANSAC algorithm and the book-by-book collation using the Gale-Shapley algorithm. We selected 3,678 book pairs with similar publication years, computed keypoint matching for 47,867,443 page pairs, and selected 411,959 page pairs by the Gale-Shapley algorithm. These values may change by the setting of various parameters.

#### vdiff.js and the Web-based Platform

The next challenge is to build a Web-based tool and platform for exploring differential reading on preprocessed results. First, we developed a JavaScript-based tool, vdiff.js<sup>4</sup>, for the page-by-page collation. As Figure 2 shows, the tool supports differential reading through direct superimposition and difference highlighting using three interfaces, such as vertical sliders and blue-red coloring, and four corresponding points set through the API or modified on the vdiff.js editor. Second, we developed a Web-based platform for differential reading<sup>5</sup> of the Bukan Complete Collection based on the results of the book-by-book collation. As Figure 3 illustrates, corresponding pages are colored by the preference score from red to green, blue to approximate the reliability of matching. The gray color represents a page without a corresponding page due to incorrect arrangement or many occurrences of editing. Here, differential reading can be extended to differential transcription, the method of transcribing only the difference, to reduce the amount of transcription along with the longitudinal publication history. Moreover, differential

---

<sup>4</sup> <http://codh.rois.ac.jp/software/vdiffjs/>

<sup>5</sup> <http://codh.rois.ac.jp/bukan/diff/>

transcription can be extended to the statistical analysis of differences, such as the activity of the publishing industry to manage the frequency of information updates.

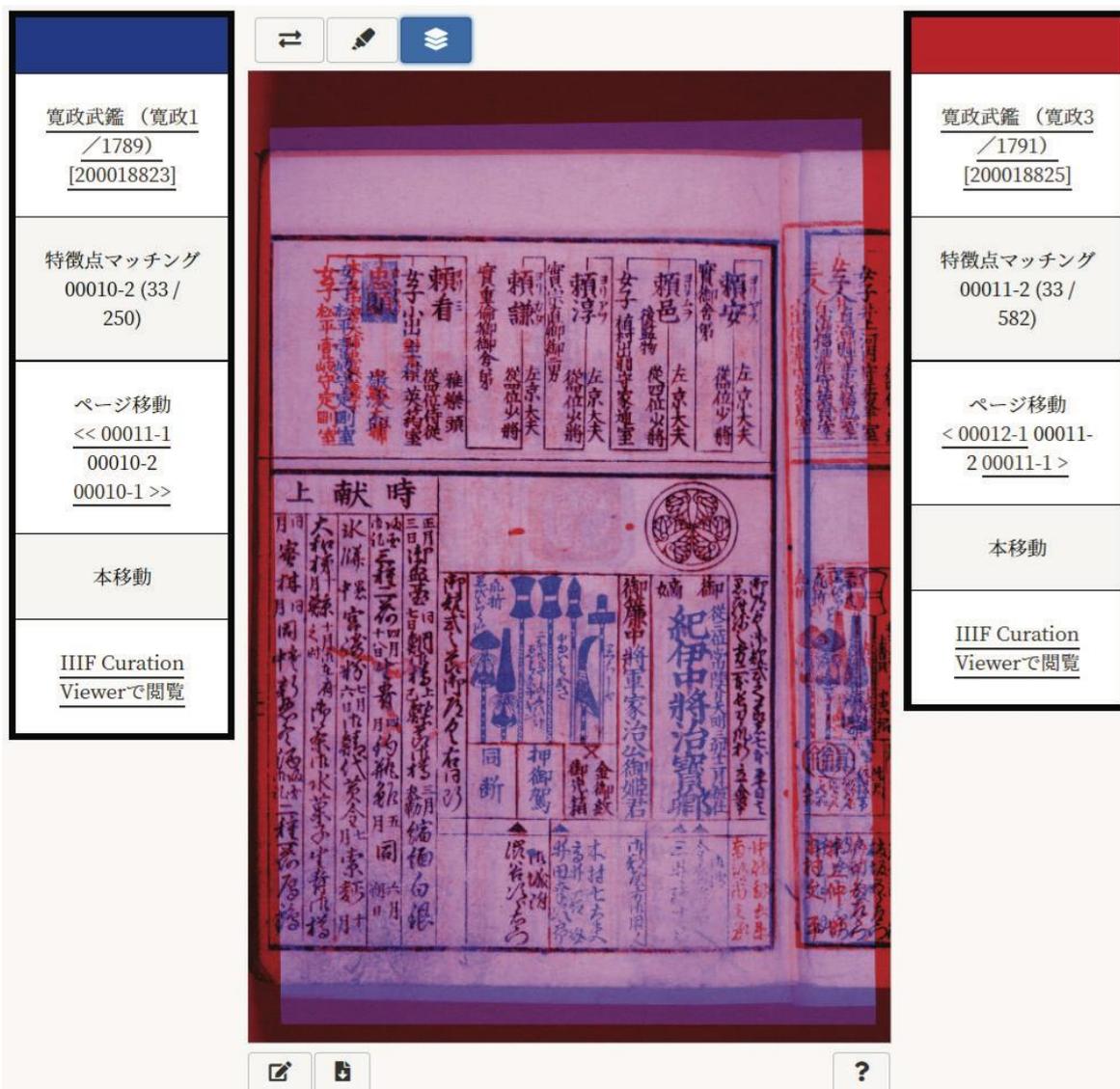


Figure 2: The result of the page-by-page collation visualized as the superimposition of two images using vdiff.js. Blue color shows pixels from the book on the left, while red color, on the right<sup>6</sup>.

<sup>6</sup> [http://codh.rois.ac.jp/cgi-bin/bukan/compare\\_page.pl?book\\_id\\_1=200018823&image\\_id\\_1=00010&book\\_id\\_2=200018825&image\\_id\\_2=00011&page\\_type=2](http://codh.rois.ac.jp/cgi-bin/bukan/compare_page.pl?book_id_1=200018823&image_id_1=00010&book_id_2=200018825&image_id_2=00011&page_type=2)

入替	寛政武鑑 (寛政1/1789) [200018823]						寛政武鑑 (寛政3/1791) [200018825]					
	00000-1	00000-2	00001-1	00001-2	00002-1	00002-2	00000-1	00000-2	00001-1	00001-2	00002-1	00002-2
ページリスト	00003-1	00003-2	00004-1	00004-2	00005-1	00005-2	00003-1	00003-2	00004-1	00004-2	00005-1	00005-2
	00006-1	00006-2	00007-1	00007-2	00008-1	00008-2	00006-1	00006-2	00007-1	00007-2	00008-1	00008-2
	00009-1	00009-2	00010-1	00010-2	00011-1	00011-2	00009-1	00009-2	00010-1	00010-2	00011-1	00011-2
	00012-1	00012-2	00013-1	00013-2	00014-1	00014-2	00012-1	00012-2	00013-1	00013-2	00014-1	00014-2
	00015-1	00015-2	00016-1	00016-2	00017-1	00017-2	00015-1	00015-2	00016-1	00016-2	00017-1	00017-2
	00018-1	00018-2	00019-1	00019-2	00020-1	00020-2	00018-1	00018-2	00019-1	00019-2	00020-1	00020-2
	00021-1	00021-2	00022-1	00022-2	00023-1	00023-2	00021-1	00021-2	00022-1	00022-2	00023-1	00023-2
	00024-1	00024-2	00025-1	00025-2	00026-1	00026-2	00024-1	00024-2	00025-1	00025-2	00026-1	00026-2
	00027-1	00027-2	00028-1	00028-2	00029-1	00029-2	00027-1	00027-2	00028-1	00028-2	00029-1	00029-2
	00030-1	00030-2	00031-1	00031-2	00032-1	00032-2	00030-1	00030-2	00031-1	00031-2	00032-1	00032-2
	00033-1	00033-2	00034-1	00034-2	00035-1	00035-2	00033-1	00033-2	00034-1	00034-2	00035-1	00035-2
	00036-1	00036-2	00037-1	00037-2	00038-1	00038-2	00036-1	00036-2	00037-1	00037-2	00038-1	00038-2
	00039-1	00039-2	00040-1	00040-2	00041-1	00041-2	00039-1	00039-2	00040-1	00040-2	00041-1	00041-2
	00042-1	00042-2	00043-1	00043-2	00044-1	00044-2	00042-1	00042-2	00043-1	00043-2	00044-1	00044-2
	00045-1	00045-2	00046-1	00046-2	00047-1	00047-2	00045-1	00045-2	00046-1	00046-2	00047-1	00047-2

**Figure 3: The result of the book-by-book collation. Background color from red to green, blue represents the ascending order of the preference score, while the gray color represents a page without a corresponding page<sup>7</sup>.**

#### 4. Conclusion

This paper proposed the book barcoding method and developed tools and platforms for differential reading of woodblock-printed books. The result is a prototype of the general-purpose visual edition comparison platform, where researchers study the evolution, or phylogenies, of editions based on visual comparison for textual and non-textual information. We expect that this will be the foundational research platform for bibliographic studies of pre-modern Japanese text.

#### Acknowledgment

The author thanks Mr. Jun Homma for his significant contribution to vdiff.js. He also thanks Prof. Kumiko Fujizane and Prof. Kazuaki Yamamoto of the National Institute of Japanese Literature for helpful comments on the research. A part of the research is based on the work of Mr. Thomas Leyh who contributed to this project while he was an NII internship student. This work is partially supported by JSPS KAKENHI Grant Number JP19H01141.

<sup>7</sup> [http://codh.rois.ac.jp/cgi-bin/bukan/select\\_page.pl?book\\_id\\_1=200018823&book\\_id\\_2=200018825](http://codh.rois.ac.jp/cgi-bin/bukan/select_page.pl?book_id_1=200018823&book_id_2=200018825)

**Reference**

- [1]. Alcantarilla, P. F., Nuevo, J., and Bartoli, A (2011) Fast explicit diffusion for accelerated features in nonlinear scale spaces. *Trans. Pattern Anal. Machine Intell*, 34:7, 1281–1298.
- [2]. Fischler, M. A., and Bolles, R. C. (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24:6, 381–395.
- [3]. Gale, D. and Shapley, L. S. (1962) College Admissions and the Stability of Marriage. *The American Mathematical Monthly*, 69:1, 9-15.
- [4]. Kitamoto, A., Horii, H., Horii, M., Suzuki, C., Yamamoto, K (2017) Structuring Time- Series Historical Sources by Human-Machine Specialization: Toward the Construction of Edo Information Platform Referring to “Bukan”. *Proceedings of IPSJ SIG Computers and the Humanities Symposium 2017*, pp. 273-280 (in Japanese).
- [5]. Kitamoto, A., Horii, H., Horii, M., Suzuki, C., Yamamoto, K. and Fujizane, K. (2018) Differential Reading by Image-based Change Detection and Prospect for Human-Machine Collaboration for Differential Transcription. *Digital Humanities 2018*.
- [6]. Leyh, T., Kitamoto, A. (2020) Computer Vision-based Comparison of Woodblock-printed Books and its Application to Japanese Pre-modern Text, Bukan. *Tenth Conference of Japanese Association for Digital Humanities (JADH2020)*, pp. 53-59.
- [7]. Moritz C, Cicero C (2004) DNA Barcoding: Promise and Pitfalls. *PLoS Biol* 2:10, e354.

## **Digital technologies and the spatial organisation of exhibitions: Interactive art as reflective experience**

**Marianna Charitonidou<sup>1</sup>**

The paper examines how augmented and virtual reality and the use of interactive digital interfaces have affected the design of exhibition spaces. The aim is to shed light on how interactive digital interfaces have influenced the way exhibition spaces are experienced. A topic that the space syntax analysis has not addressed comprehensively is the impact of interactive technologies on how the visitors experience exhibition spaces. The paper also explains why the concept of “spatial configuration”, which is central for the space syntax approach, is pivotal for better grasping the relationship between new media art and the architecture of exhibition spaces, and their respective use of augmented and virtual reality. It explores an ensemble of immersive art examples, paying special attention to the distinction between immersion and interactivity. Its objective is to render explicit how extended reality technologies contributed to the design of immersive experiences, influencing significantly the interrelations between the technical, aesthetic and institutional aspects not only of exhibition design, but of the dissemination of artworks as well.

At the centre of the paper is the role of extended reality technologies in designing immersive experiences in the case of art practices that place particular emphasis on participation, interaction, technology and digital media. A new kind of subjectivity emerges thanks to the development of the so-called immersive art. Panayiota A. Demetriou, in “Imagineering’ mixed reality (MR) immersive experiences in the postdigital revolution: innovation, collectivity, participation and ethics in staging experiments as performances’, used the term ‘imagineer’ to describe this new kind of subjectivity. According to Demetriou, “[t]he Imagineer is an interaction designer, an experience designer, a user experience researcher, a facilitator, a connector and networker, a translator, a project manager, a visionary entrepreneur” (2018, p. 170). The shifts in subjectivity concern both the visitor and the creator. The visitors adopt a more active role, which is achieved thanks to their interactivity or interaction with the artworks. In parallel, the status of the creators is transformed significantly in the case of experiential immersive art. The artworks are not any more related to the intentionality of the artist that conceives them and leads the process of their making but are the result of a much more complex and transdisciplinary process, which can be achieved thanks to the formation of multidisciplinary art collectives such as the teamLab. Symptomatic of this stance is the fact that the people that work for teamLab use the term ‘ultra-technologist’ to describe their

---

<sup>1</sup> ETH Zurich

professional activity or discipline or field of expertise and not the most conventional term ‘artist’. The shifts that take place in the field of arts, curation and museums do not concern only the artists and the visitors but the whole system of dissemination and promotion of the arts, including all its institutional aspects. The system of financing the artists and the museums is transform as well as the status of art galleries.

An important distinction is that between interactivity and immersion. According to Panayiota A. Demetriou, interactivity and immersion differ in the sense that the former involves “attentiveness to signs”, while the latter “occurs at the disappearance of signs” (2018, p. 177). Useful for understanding when an experience is immersive is the remark that “for an experience to be considered immersive it must be more than a three-dimensional image that surrounds a user”. The current trends in immersive art are characterized by the tendency to prioritize augmented reality instead of virtual reality. A common critique of virtual reality is related to the fact that it – virtual reality – “has been considered to restrict immersion by isolating its users, not only the person wearing the headset, but also anyone standing near them”. (ibid., p. 178)

The coexistence of the virtual and the physical enhances the sense of immersion, and the interaction not only between the visitor and the artwork, but also that between the visitors. To fully grasp the transformations that immersive art provokes, we should seriously take into consideration the interrelations between the technical, the artistic, and the institutional aspects. Augmented reality is just one of the various forms of mixed reality technologies that can be used in exhibition design and in the creation of immersive art artworks. The feature of augmented reality that is at the centre of this paper is the coexistence of the digital content and the physical world. To have an overview of the generalized use of virtual reality and augmented reality technologies in museums and art galleries in Europe, we can bring to mind that, in 2015, they “piloted in exhibitions in over a quarter of European museums”, as Richard Yu-Chang Li and Alan Wee-Chung Liew underscore, in ‘An interactive user interface prototype design for enhancing on-site museum and art gallery experience through digital technology’ (2015). This was the case in 2015 and today the use of virtual reality and augmented reality technologies in museums and art galleries in Europe is much more generalized than back then.

### **Selective References**

Demetriou, Panayiota A. (2018) ‘Imagineering’ mixed reality (MR) immersive experiences in the postdigital revolution: innovation, collectivity, participation and ethics in staging experiments as performances’, *International Journal of Performance Arts and Digital Media*, 14(2), pp. 169-186.

- Navarrete, T. (2019) 'Digital heritage tourism: innovations in museums', *World Leisure Journal* 61(3), pp. 200-214. doi: 10.1111/muse.12201
- Negroponte, N. (1995) *Being Digital*. New York: Alfred A. Knopf.
- Parker, E. and Saker, M. (2020) 'Art museums and the incorporation of virtual reality: Examining the impact of VR on spatial and social norms', *Convergence*, 26(5–6), pp. 1159–1173. doi: 10.1177/1354856519897251.
- Paul, C., ed. (2016) *A Companion to Digital Art*. London: Wiley-Blackwell.
- Penn, A. (2003) 'Space Syntax and Spatial Cognition: Or Why the Axial Line?', *Environment and Behavior*, 35(1), pp. 30–65. doi: 10.1177/0013916502238864.
- Shanken, E. A. (2016) 'Contemporary Art and New Media Digital Divide or Hybrid Discourse?', in Paul, C., ed. (2016) *A Companion to Digital Art*. New York: Wiley-Blackwell, pp.463-481.
- Shehade, M. and Stylianou-Lambert, T. (2020) 'Virtual Reality in Museums: Exploring the Experiences of Museum Professionals', *Applied Sciences*, 10(11), p. 4031. doi: 10.3390/app10114031
- Simanowski, R. (2011) *Digital Art and Meaning: Reading Kinetic Poetry, Text Machines, Mapping Art, and Interactive Installations*. Minneapolis: University of Minnesota Press.
- vom Lehn, D., Heath, C. and Hindmarsh, J. (2001) 'Exhibiting Interaction: conduct and collaboration in museums and galleries', *Symbolic Interaction*, 24, pp. 189–216.
- Westerby G. and Keegan, K. (2019) 'Digital Art History and the Museum: The Online Scholarly Collection Catalogues at the Art Institute of Chicago', *Visual Resources*, 35(1-2), pp. 141-154, doi: 10.1080/01973762.2018.1553445
- Wang, X. (2009) 'Augmented Reality in Architecture and Design: Potentials and Challenges for Application', *International Journal of Architectural Computing*, 7(2), pp. 309–326. doi: 10.1260/147807709788921985.
- Wineman, J. D. and Peponis, J. (2010) 'Constructing Spatial Meaning: Spatial Affordances in Museum Design', *Environment and Behavior*, 42(1), pp. 86–109. doi: 10.1177/0013916509335534.
- White, M. (1997) 'Cabinet of Curiosity: Finding the Viewer in a Virtual Museum', *Convergence*, 3(3), pp. 29–71. doi: 10.1177/135485659700300305.
- Yu-Chang Li, R. and Wee-Chung Liew Alan (2015) 'An interactive user interface prototype design for enhancing on-site museum and art gallery experience through digital technology', *Museum Management and Curatorship*, 30(3), pp. 208-229. doi: 10.1080/09647775.2015.1042509



This paper analyses the decentering and subsequent *re*-centering of language in Hanafi's *A Dictionary of the Revolution*. It views *A Dictionary of the Revolution* as an example of experimental decentered text similar to Jacques Derrida's *Glas* (1974). Following Derrida's approach as applied to digital literature/language art as proposed by John Cayley (2018), it also views language as "media-agnostic". In addition to the digital work's form, it also examines the work's code through Mark Marino's *Critical Code Studies* (2020). Marino writes:

But the code is not enough in itself. It is crucial to explore context. Who wrote the code? When and why? In what language was the code written? What programming paradigm was used? Was it written for a particular platform (hardware or software)? How did the code change over time? What material or social constraints impacted the creation of this code? How does this code respond to the context in which it was created? How was the code received by others? (28)

To explore the architecture and impact of this digital structure, this work is presented as a born-digital essay that engages directly with the code used to format and structure *A Dictionary of the Revolution* (permission to use this code developed by Youssef Faltas for this purpose has been granted by Amira Hanafi). This research is regarded as 'practice-led research' as defined by Smith and Dean (2009). This paper also extends current digital literary review formats proposed in the publication *The Digital Review*.

This paper not only reflects on the digital structure, but directly engages with it. As born-digital works utilize digital code and technology to create new meaning, it therefore follows that criticism should also utilize the same digital code and technology. In so doing, this practice-led research develops a better understanding of Hanafi's form, that in turn can be proposed as a new interconnected digital essay format that can be classified as a 'third-generation electronic literature' form (as defined by Flores 2019). Such a form, I argue, could be utilized to better represent the collective, polyphonic response to major political and global events, such as a response to the COVID-19 pandemic by depicting decentered, kaleidoscopic viewpoints.

## References

- [1]. Ackermans, H. (2019) *ELMCIP*. 'A Dictionary of the Revolution'. Available at <https://elmcip.net/creative-work/dictionary-revolution> [accessed on 18 June 2021]
- [2]. Cayley, J. (2018) *Grammalepsy: Essays on Digital Language Art*, London: Bloomsbury Academic.

- [3]. Derrida, J. (1978) [trans. A. Bass] *Writing and Difference*, London: Routledge.
- [4]. Derrida, J. (1986) *Glas*, Lincoln: University of Nebraska Press.
- [5]. Flores, L. (2019) 'Third Generation Electronic Literature', *Electronic Book Review*, 7 April, at <https://electronicbookreview.com/essay/third-generation-electronic-literature/> (accessed 4 May 2021).
- [6]. Hanafi, A. (2016) *A Dictionary of the Revolution*. Available at <https://www.ibraaz.org/projects/143> [accessed on 18 June 2021]
- [7]. Hanafi, A. (2017) *A Dictionary of the Revolution*. Available at <http://www.qamosalthawra.com/en> [accessed on 18 June 2021]
- [8]. Marino, M. (2020) *Critical Code Studies*, Cambridge: MIT Press.
- [9]. Smith, H. and Dean, R.T. (2009) *Practice-led Research, Research-led Practice in the Creative Arts*. Edinburgh: Edinburgh University Press.

# Experimental LDA Topic Modelling of Tennyson's Epic Poems

Iku FUJITA<sup>1</sup>

## Introduction

This study aims to investigate the application of Latent Dirichlet Allocation (LDA), a type of topic model [1], on verse texts and its effectiveness in poetry research. Topic modelling is considered a promising approach in the field of digital humanities and text mining [2]. While Tabata [3], Kiyama [4], Huang [5] and some other studies have examined prose texts using topic modelling, few studies, apart from Rhody [6], Navarro-Colorado [7], Henrichs [8], and Okabe [9], have investigated the possibility of applying topic modelling to poetry. The use of the method still has room for improvement with regard to its application to a corpus of poems. This paper purports to report emerging results of running LDA on a corpus of verse texts and discuss the feasibility of practicing LDA with small size segments, using the Victorian poet Alfred Tennyson's poems.

## Methodology

When LDA is employed on a prose text, the texts are often divided into consecutive segments of equal size. Some set the segment size to 1,000 words [10], 2,000 or greater [11], depending on their research questions. One of the most crucial decisions to make in employing LDA on verse texts is the size of a segment. As shown in Table 1, the vast majority of Tennyson's poems, his lyrical poems in particular tend to be shorter than 1,000 words apiece, while the poet wrote not a few epic poems, which tend to have a larger number of word tokens in comparison with lyrical poems. Thus, the segment size should be set carefully depending on the corpus in use. This study uses his 26 epic poems which consist of more than 2,000 words so that each work is divided into at least two segments with 1,000 words each (Table 2).

Masada [12] reports that LDA is successfully applied on a short text corpus (approximately ten words) as long as hyperparameters and other preferences are appropriately adjusted. However, since his target corpus is the titles of English journals, it is questionable whether his observations hold good for a short prose/verse text corpus. In this study, we used three different settings: slicing texts into 1,000-, 100-, and 25-word consecutive segments, to examine whether a small segment size is also applicable to poetry corpus. Each setting was decided on the basis of the composition of the corpus in terms of the size of each of the poems included. As shown in Table 1, the group of poems with 100–199 word-tokens is the largest group in the set; thus, 100 was adopted as a tentative

---

<sup>1</sup> Student at Graduate School of Language and Culture, University of Osaka

standard of segment size. Additionally, the shortest poem has 26 tokens; therefore, the most approximate value 25 was set as the minimum of segment size. When each text was sliced into equal-sized consecutive segments, final two parts were joined together unless the final chunk was exactly 1,000, 100 or 25 words in length.

Before running the LDA, function words and proper nouns were excluded as stopwords. On the other hand, all adverbs are included in the examination. Table 3 shows the conditions of LDA topic modelling of each segment size, and all conditions are set to the same to see whether the segment size differences influence the topic modelling results.

**Table 1: The descriptive statistics of Tennyson's Poetical Works.**

The number of poems		424
Total tokens		342,073
Shortest poem in num. of words		26
Longest poem in num. of words		26,749
Mean tokens per poem		806.78
Standard deviation		2,172.00
	$1 \leq n < 100$	80
	$100 \leq n < 200$	116
	$200 \leq n < 300$	67
	$300 \leq n < 400$	26
	$400 \leq n < 500$	20
	$500 \leq n < 600$	16
The No. of poems with total	$600 \leq n < 700$	13
tokens of	$700 \leq n < 800$	8
	$800 \leq n < 900$	7
	$900 \leq n < 1,000$	4
	$1,000 \leq n < 5,000$	51
	$5,000 \leq n < 10,000$	11
	$n \leq 10,000$	5

**Table 2: The list of 26 epic poems of Tennyson.**

No.	Year of Publication	Title of Poems	The Num. of Tokens
1	1847	Princess	26,772
2	1859	Lancelot and Elaine-Idylls of the King	11,959
3	1872	THE ROUND TABLE Gareth and Lynette-Idylls of the King	10,851
4	1855	Maud	10,280
5	1864	Aylmers Field	9,461
6	1857	Geraint and Enid-Idylls of the King	8,035
7	1857	Merlin and Vivien-Idylls of the King	8,024
8	1869	The Holy Grail-Idylls of the King	7,663
9	1862	Enoch Arden	7,531
10	1857	The Marriage of Geraint-Idylls of the King	6,951
11	1871	The Last Tournament-Idylls of the King	6,298
12	1859	Guinevere-Idylls of the King	5,802
13	1885	Balin and Balan-Idylls of the King	5,090
14	1869	Pelleas and Ettarre-Idylls of the King	5,043
15	1869	The Coming of Arthur-Idylls of the King	4,313
16	1869	The Passing of Arthur-Idylls of the King	3,889
17	1889	The Ring	3,752
18	1886	Locksley Hall Sixty Years After	3,355
19	1842	Morte dArthur	2,519
20	1880	The Sisters	2,357
21	1842	Locksley Hall	2,326
22	1842	The Gardeners Daughter	2,282
23	1868	Lucretius	2,238
24	1842	OEnone	2,090
25	1832	The Palace of Art	2,069
26	1880	Columbus	2,007

**Table 3: The conditions of LDA topic modelling.**

	Seg. Size 1,000	Seg. Size 100	Seg. Size 25
No. of topics		10	
No. of iterations		1,000	
Optimize interval		20	
Optimize burn in		50	

**Results**

Tables 4 to 6 show the list of ten topics with their key words when the segment size was set to 1,000, 100, and 25, respectively. The label of each topic was represented by three most contributing words to the topic. All and then appeared in multiple topics in three Tables. Moreover, vocatives and titles of the characters, as king, sir, lord, and so forth, influenced the results, as topics 1, 3, 6, 7 (Table 4). As Kuroda [11] points out, deciding what constitutes stopwords is a ticklish issue for topic modelling. How to decide stopwords avoiding arbitrariness should be discussed in a different paper.

**Table 4: The list of topics with their key words (seg. Size 1,000).**

Topic No.	0	1	2	3	4	5	6	7	8	9
Alpha value	0.202	0.116	0.122	0.165	0.146	0.210	0.229	0.569	3.126	0.178
Keys	alone	lady	earth	prince	ring	fame	king	king	all	holy
	happy	princess	hall	earl	father	charm	sir	sir	then	sin
	die	woman	spain	rode	home	storm	deep	knight	love	hall
	soul	prince	grave	court	house	master	spake	queen	said	saw
	gods	florian	dead	arms	poor	mood	great	knights	man	quest
	mother	women	garden	eat	seem'd	boon	water	then	yet	crown
	hear	men	forward	hall	girl	woods	answer	horse	now	grail
	harken	read	nature	dress	gone	use	moon	answered	came	lord
	dying	head	glory	gay	children	pretty	sun	shield	made	galahad
	seem'd	laws	god	hawk	marriage	careless	brand	damsel	heart	brother
	law	college	race	sparrow	grave	cause	drew	spake	here	heaven
	tears	south	moor	doorm	look'd	sole	rain	being	let	vow
	knowledge	girls	divine	lord	year	wise	bold	field	ever	vision
	golden	highness	war	fight	babe	snake	barge	lady	men	seen
	hills	something	madness	looked	ask'd	yield	hast	noble	hand	madness
	grow	boys	brain	fall	mother	eagle	round	son	day	walls
	cheek	talked	beat	armour	turn'd	mountain	heathen	fool	great	royal
	paris	soldier	locksley	noble	answer'd	rhyme	wind	table	old	earth
	heavy	swallow	chains	town	long	smiling	throne	knave	eyes	bors
	foot	answered	kind	wine	t'amo	tame	black	hast	life	sware

**Table 5: The list of topics with their key words (seg. Size 100).**

Topic No.	0	1	2	3	4	5	6	7	8	9
Alpha value	0.12073	0.11591	0.33551	0.13202	0.7533	0.14849	0.26948	0.30264	1.50888	0.26256
Keys	fool	men	king	mere	all	ring	light	knight	all	all
	old	princess	sir	sleep	then	home	rose	sir	love	long
	forward	every	all	brand	came	father	sun	then	man	fire
	earth	read	holy	arm	saw	mother	flowers	prince	said	heaven
	spain	lady	then	moon	stood	house	air	king	yet	sea
	turn	florian	knights	sword	hand	year	night	queen	then	day
	age	laws	great	lightly	face	babe	alone	said	now	great
	war	win	spake	drew	hall	children	west	cried	heart	city
	swine	knowledge	made	white	rose	often	wild	horse	let	far
	hell	college	queen	saying	fell	seem'd	hear	damssel	life	land
	lower	woman	round	nest	night	gone	garden	lord	well	ever
	dagonet	silken	table	hilt	eyes	death	here	answered	sweet	deep
	moor	highness	hast	breath	past	cared	die	fair	world	earth
	chains	something	old	lake	sat	ghost	deep	knave	know	saw
	wheel	ruin	knight	quick	dark	day	rain	shield	ever	built
	race	sang	kings	harm	dead	abroad	fair	here	too	shore
	mind	time	people	caught	arms	kindly	morning	quest	men	work
	divine	wise	realm	sir	light	turn'd	far	ride	come	low
	locksley	wild	lord	ridge	old	poor	mother	lady	made	shadow
	crowd	strange	see	winter	cry	oft	happy	hall	child	flame

**Table 6: The list of topics with their key words (seg. Size 25).**

Topic No.	0	1	2	3	4	5	6	7	8	9
Alpha value	0.5404	0.20157	0.09905	0.08963	0.11062	0.07479	0.72813	0.04854	0.43749	0.22649
Keys	king	all	war	turn	horse	ran	all	tall	then	all
	all	light	men	fool	left	gold	love	every	all	sea
	then	sun	old	nature	right	hawk	yet	fairy	came	night
	said	star	realm	work	arms	sparrow	man	larger	hand	land
	sir	red	king	man	sword	jewels	heart	scorn	saw	wind
	knight	white	land	song	fallen	bird	life	tarn	rose	far
	here	fire	heathen	fine	huge	together	sweet	takes	stood	wild
	lord	rose	place	law	struck	dragon	mother	headed	eyes	ever
	made	heaven	kings	swine	horses	cup	ever	debt	past	long
	queen	eyes	iron	fail	song	college	now	text	face	day
	name	flowers	spain	hands	armour	making	well	gray	found	heard
	fair	morning	common	lower	strong	blaze	let	shook	rode	world
	men	round	traitor	false	fierce	hour	dead	watch	hands	voice
	good	garden	christ	beast	spring	finger	know	circle	head	dark
	answered	blood	lords	science	lance	careless	child	ancient	sat	storm
	great	air	people	powers	naked	drank	said	female	fell	wood
	knights	summer	rome	wheel	leapt	catch	death	tallest	king	great
	cried	golden	flash	higher	point	silken	then	marge	took	full
	spake	moon	chains	forward	shield	gilded	too	ample	again	low
	prince	cloud	bones	play	lips	quiet	father	creep	now	hall

The most prevalent topic (according to the hyperparameter  $\alpha$ ) in each of the three LDA runs is shown with its top 20 key words and word-weight details (Table 7). Table 8 shows the  $\alpha$  values of all the topics. Correlation coefficients between the topics with the highest  $\alpha$  value are given in Table 9. The correlation coefficients were calculated based on word-weights. Table 9 suggests the correlation coefficients tend to get lower when the ratio of segment size difference is larger. However, the correlation coefficients, as a whole, indicated high positive correlations; therefore LDA can also be applied to the small segment size of poetry.

**Table 7: Keys and weights of three highest  $\alpha$  value topics.**

Segment size: 1,000 Topic 8		Segment size: 100 Topic 8		Segment size: 25 Topic 6	
Alpha value	3.13	Alpha value	1.51	Alpha value	0.73
Keys	Weights	Keys	Weights	Keys	Weights
all	1289.10	all	727.08	all	439.08
then	693.10	love	421.08	love	419.08
love	422.10	man	333.08	yet	336.08
said	414.10	said	328.08	man	265.08
man	351.10	yet	310.08	heart	264.08
yet	347.10	then	307.08	life	212.08
now	313.10	now	294.08	sweet	185.08
came	300.10	heart	269.08	mother	183.08
made	284.10	let	257.08	ever	178.08
heart	266.10	life	203.08	now	178.08
here	258.10	well	200.08	well	171.08
let	256.10	sweet	190.08	let	171.08
ever	241.10	know	188.08	dead	162.08
men	226.10	world	188.08	know	155.08
hand	225.10	ever	183.08	child	153.08
day	222.10	too	175.08	said	149.08
great	221.10	men	162.08	death	146.08
old	221.10	come	157.08	then	144.08
eyes	217.10	made	152.08	too	142.08
life	212.10	child	151.08	father	141.08

**Table 8:  $\alpha$  values of 10 topics, three segment sizes.**

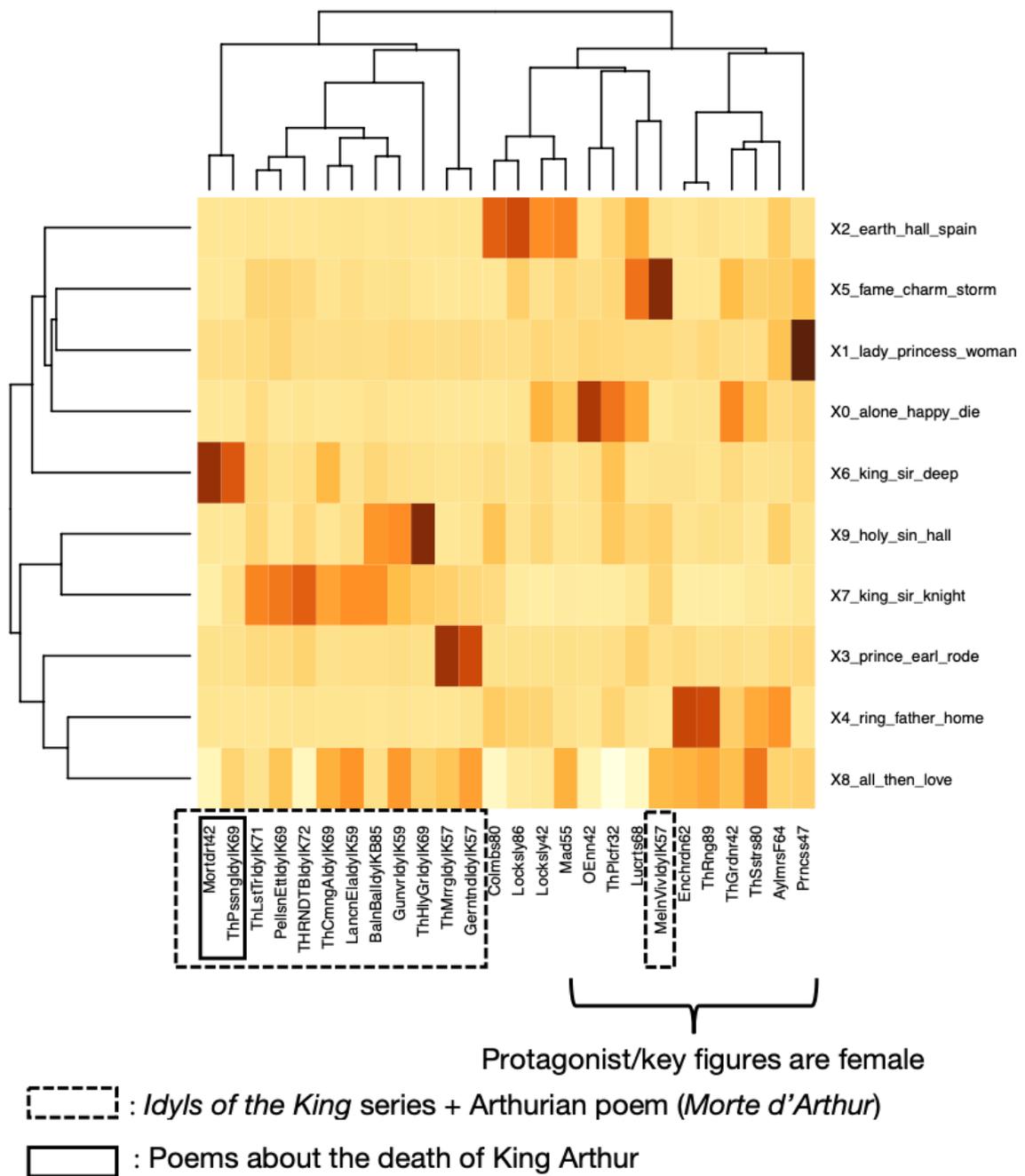
	topic 0	topic 1	topic 2	topic 3	topic 4	topic 5	topic 6	topic 7	topic 8	topic 9
segment size: 1,000	0.202	0.116	0.122	0.165	0.146	0.210	0.229	0.569	3.126	0.178
segment size: 100	0.121	0.116	0.336	0.132	0.753	0.148	0.269	0.303	1.509	0.263
segment size: 25	0.540	0.202	0.099	0.090	0.111	0.075	0.728	0.049	0.437	0.226

**Table 9: Correlation of word weights (The highest  $\alpha$  value topic of each examination)**

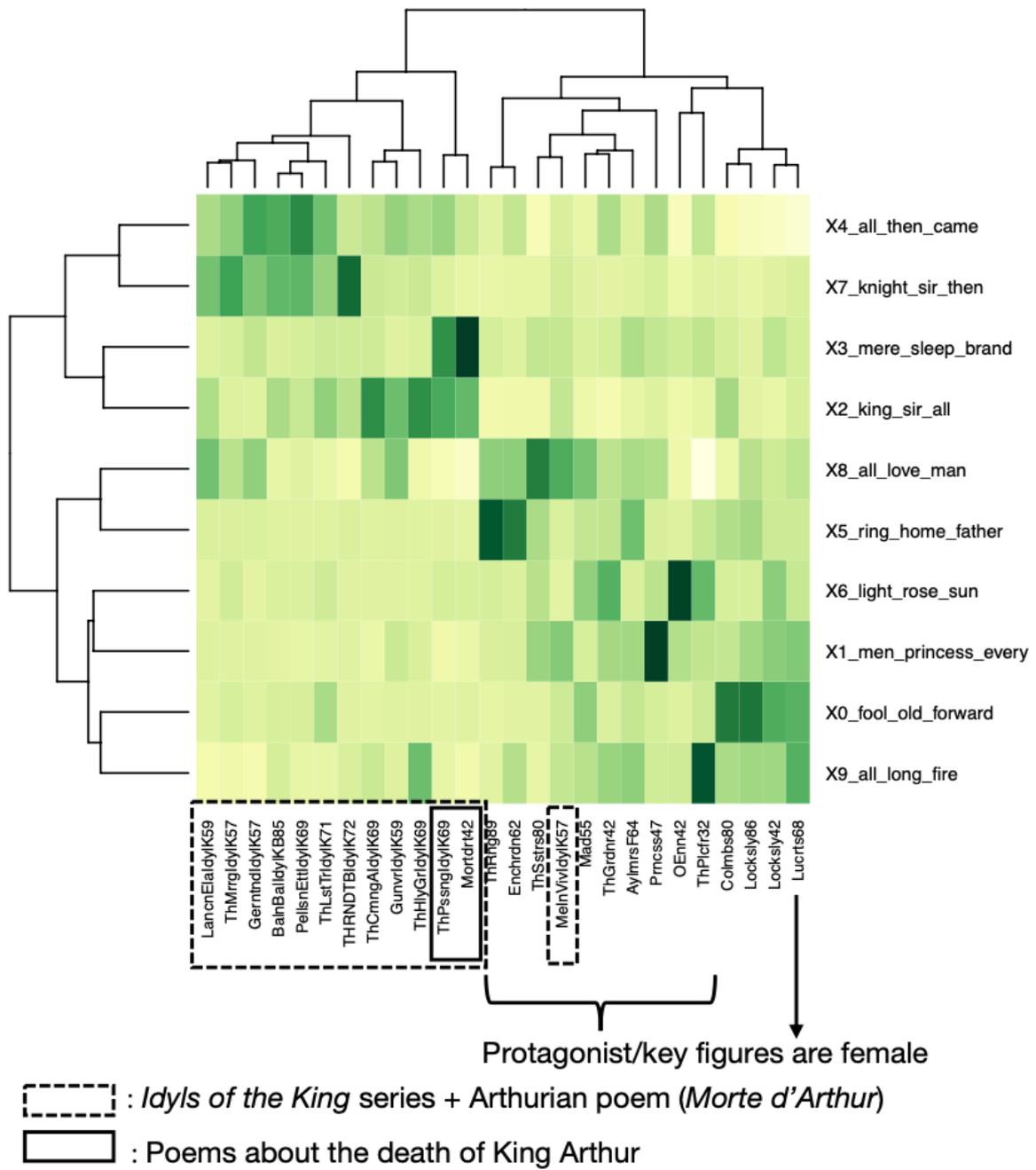
	segment size: 1,000 Topic 8	segment size: 100 Topic 8	segment size:25 Topic 6
segment size: 1,000 Topic 8	-	0.881	0.785
segment size: 100 Topic 8		-	0.914
segment size:25 Topic 6			-

Figures 1, 2, and 3 show heatmap representations of the outputs of LDA, segment size 1,000, 100, and 25, respectively. In each Figure, ten most prominent topics are arrayed vertically, and the 26 poems are arrayed horizontally. In three Figures, the *Idyls of the King* series and an Arthurian poem lie in the same major cluster, and the other non-*Idyls* works belong to the right major cluster. In Figure 1 and 2, a poem from the *Idyls* series, *Merlin and Vivien* (MelnVivIdylK57), is located in non-*idyls*' cluster, amongst the works in which females take the essential roles as key figures because topic 5 tends to be shown as a prominent topic of MelnVivIdylK57 in Figure 1, while topic 7 is significant in *Idyls* cluster.

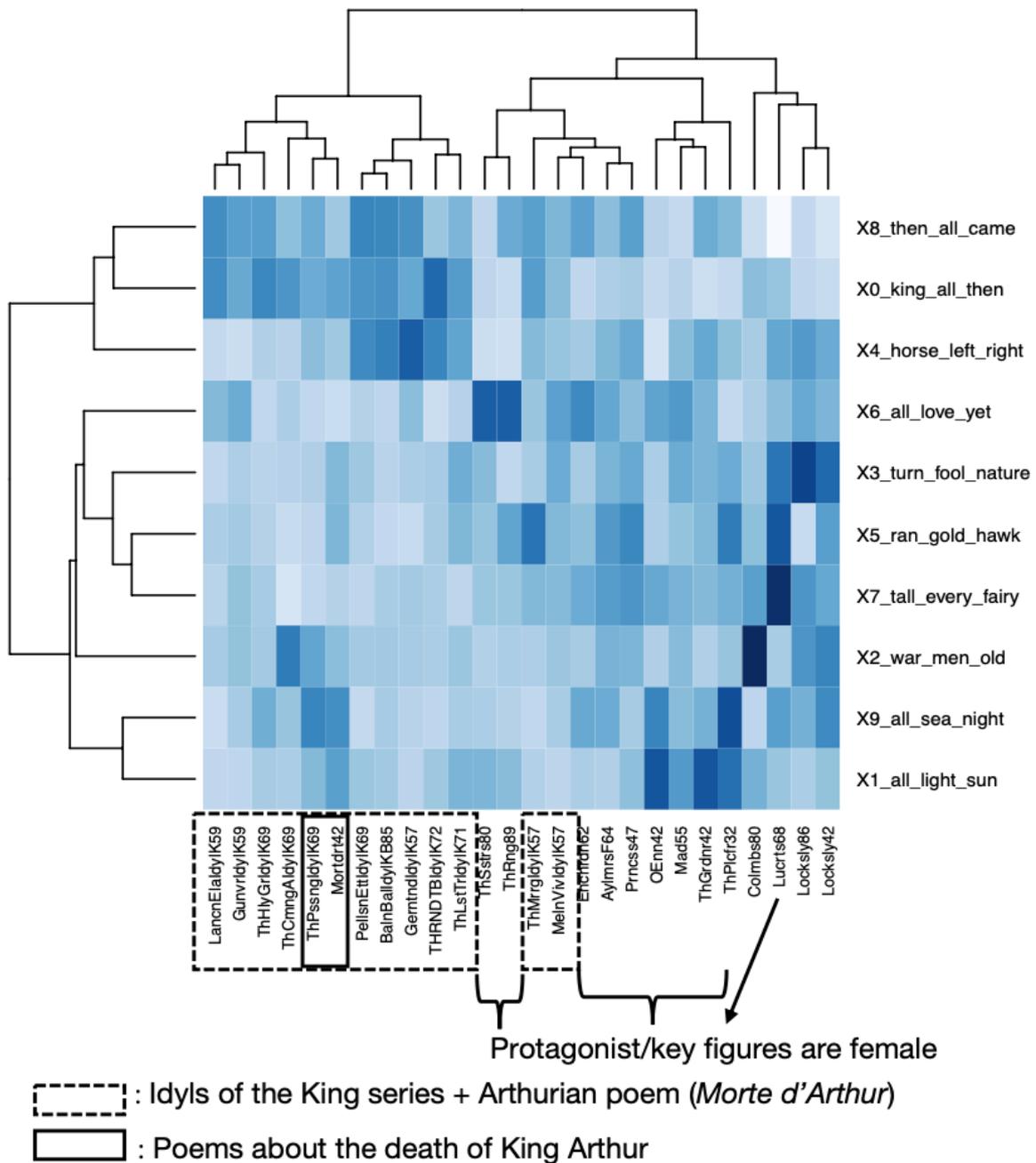
From Figure 1 to Figure 3, although there can be seen subtle differences, the clusters are remarkably similar. The *Idyls* series tend to be located in the same major cluster, and others are in the other major cluster.



**Figure 1: The heat map of LDA topic modelling result (seg. size: 1,000; No. of topic: 10).**



**Figure 2: The heat map of LDA topic modelling result (seg. size: 100; No. of topic: 10).**



**Figure 3: The heat map of LDA topic modelling result (seg. size: 25; No. of topic: 10).**

**Implications for Future Study**

To sum up, the findings of this study indicate that LDA is an adequate and plausible method for poetry study as well as prose text study. However, future investigations are necessary to validate the conclusions that can be drawn from this study, especially observing other diagnostics values and finding the most relevant way to presume the number of topics.

**Reference**

- [1]. T Blei, M. D., Ng, Y. A., and Jordan, I. M. (2003). "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3, 2003, pp. 993-1022.
- [2]. Meeks, E. and Weingart, B. S. (2012). "The Digital Humanities Contribution to Topic Modeling." *Journal of Digital Humanities*, vol.2, No. 1 Winter 2012, pp. 1-6.
- [3]. Tabata, T. (2017). "Mapping Dickens's Novels in a Network of Words, Topics, and Texts: Topic Modelling a Corpus of Classic Fiction." *Japanese Association for Digital Humanities Conference 2017*, September 2017, Doshisha University.
- [4]. Kiyama, N. (2018). "How Have Political Interests of U.S. Presidents Changed?: A Diachronic Investigation of the State of the Union Addresses through Topic Modeling." *English Corpus Studies*, Vol. 25, pp. 79-99.
- [5]. Huang, C. (2020). "Quantitative Analysis of Chinese Mystery Novels: Focusing on the Works of Cheng Xiao Qing and Gui Ma Xing." *Studies in Language and Culture Osaka University*. 2019, pp. 31-45.
- [6]. Rhody, M.L. (2012). "Topic Modeling and Figurative Language." *CUNY Academic Works*, 2012, pp. 19-35.
- [7]. Navarro-Colorado, B. (2018). "On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of Spanish Poetry." *frontiers in Digital Humanities*, 20, 2018, pp. 5-15.
- [8]. Henrichs, A. (2019). "Deforming Shakespeare's Sonnets: Topic Models as PoemsAuthor(s)." *Criticism*, Vol. 61, No. 3, pp. 387-412.
- [9]. Okabe, M. (2019). "Thou and You in Emily Dickinson's Poems Using Topic Modeling: Reconsideration of Interjections." *Proceedings of Japanese Association for Digital Humanities Conference 2019*, pp. 125-131.
- [10]. Huang, C. (2020). "Experimental Topic Analysis on Chinese Mystery Prose Texts." *Studies in Language and Culture Osaka University*. 2020, pp. 1-17.
- [11]. Kuroda, A. (2017). "Quantitative Analysis of Literary Works: Novels of Sir Arthur Conan Doyle." *Studies in Language and Culture, Osaka University*. 2016, pp. 23-41.
- [12]. Masada, T. (2019). "Topic model-no Kiso-to Ouyou (The Basis and Application of Topic modelling)." *Japanese Society of Computational Statistics Study Group "IR (Institutional Research)-no tame-no Toukeiteki Model Kouchiku-ni Kansuru Kenkyuu (Research on the Statistical Model Construction for the Institutional Research)" Workshop*, March 2019, The Institute of Statistical Mathematics.

# A study on the readerly aspects of Electronic Poetry through Cognitive Poetics

Mariyam Nancy J<sup>1</sup>, David Arputharaj<sup>1</sup>

## Introduction

Digital literature or electronic literature takes its roots from concrete, visual and Dadaist poetry due to its similarity in composition and form. Digital literature is characterised by poly vocal expressions, reader participation, and nomadic nature of the text and the presence of audio-visual elements. This study deals with digital poetry, a genre of digital literature, which is composed with the aid of programming languages, such as JAVA and HTML, and other media devices, like Flash. The emergence of digital literature demanded new theoretical frameworks that explored textuality, composition, reading and comprehension.

Literary reading is a two-fold process that consists of perceiving the symbols offered by the text and of making mental representations. The former is involved with the dissemination of signs that are present in the text; the other is all about meaning making and mental representation. Theories concerning digital literature serve as an interpretative framework, through which particular instances of the text can be read. Hypertext theorists have addressed the concerns of the non-linear reading, multi-modality and interactivity and have endorsed on close reading based on semiotics, linguistics and narratology. Cognitive turn in hypertext theories was necessitated due to the resemblance between the working of human cognition and hypertext literature as both worked on “assosiativeness” that existed between the sign and meaning.

## Research Approach

This research has taken three major components of literature for exploring the literariness of digital literature, namely text, textuality and text world. Roland Barthes, in his seminal essay “From Work to Text,” defines the text as a plural, active and imaginary continuum that is experienced only in activity or production and this resonates with the features of digital poetry. Textuality is the coherence of form; symbols and meaning that provide accessibility of the text to the reader. A text world is the mental representation that is created from a combination of the spatial and the temporal cues offered through a narrative that allows the reader to imagine and inscribe them in memory. Therefore, the act of reading is more cognitive than linguistic. Such an undertaking requires an inquiry that

---

<sup>1</sup> Bharathiar University

differentiates the cognitive implications of reading a text produced in print as well as a digital medium.

This study aims to approach the readerly aspects of digital poetry through cognitive poetics, as this area deals with the implications of the formal aspects of literary texts on cognitive activities. Cognitive Poetics specializes in addressing the nuances of a text from a perceptual and creative level. This research aims to address elements such as foregrounding, prototypicality, attention and deixis that are involved in anchoring the idea of a text. It also endorses on Text World Theory to understand the meaning making process while reading a work of electronic literature

### **Conclusion**

The conception of digital literature begins with the ontological understanding of this literary form and this research is an attempt to do so. The digital medium has become an indispensable part of human existence and the influence it exerts on poetic practices is gaining momentum. Digitalisation of literature, composition, medium or otherwise is not an invasion of traditional literary spaces but a renaissance through a new medium, empowered by the aesthetics of the digital that is open to all of mankind.

### **Reference**

- [1]. White, A. (2007). "Understanding Hypertext Cognition: Developing Mental Models to aid user's comprehension." *First Monday*, 12(1).
- [2]. Memmot, T. (2011). "Beyond Taxonomy: Digital poetics and the Problem of Reading." *Digital Rhetoric and Poetics: Signifying Strategies in Electronic Literature*, pp. 57-73.
- [3]. Stockwell, P. (2002). *Cognitive Poetics: An Introduction*. New York: Bloomsbury Press.

# **Architectural Drawings Exposed and the Effect of Digitization: The Rise of Artefactual Value vs the Democratization of Knowledge**

**Marianna Charitonidou<sup>1</sup>**

Architecture exhibitions are vehicles of architectural knowledge dissemination and constitute sites of methodological innovation. Two phenomena, pivotal for comprehending the place of drawings in architecture exhibitions, are examined conjointly in this paper: the first is the rise of architectural drawings' artefactual value, triggered by a series of exhibitions at the Max Protetch, Leo Castelli and Rosa Esman galleries. The second phenomenon is the effect of the digitization of architectural drawings. In-situ research on drawings is part of the institutions' policies for the vitalization of debates. The democratization accompanying digitization causes an increase in the fascination caused by the aura of the access to the original. Michel Foucault's conviction that knowledge, power and subjectivity are by no means contours given once and for all, but indeed series of variables which supplant one another, is a starting point for understanding the complexity of the double agency of democratizing and fetishizing architectural drawings. Useful for understanding the tension between the democratization of knowledge that accompanies the digitization of architectural drawings and the intensification of their fetishization is Michel Foucault's understanding of the notion of knowledge and the distinction he draws between the concept of *connaissance* and the concept of *savoir*. For Foucault, *connaissance* means knowledge in the sense of the subject's relationship to an object and the rules which govern this relationship, while *savoir* means knowledge in the sense of the underlying structure which is the precondition of any *connaissance*. Christian Fuchs, in *Marx in the Age of Digital Capitalism*, distinguishes two broad approaches in the so-called Internet Studies: on the one hand, an approach that is based on a cultural studies background, and, on the other hand, an approach that is based on a political economy background. Fuchs claims that the theoretical background of the first is post-structuralist, while that of the second is Marxist. The paper, taking as its starting point the fact that capitalism is based on the production of technological means, which are by no means neutral, explores the relationship between the effect of digitization of the special collections of architecture and art museums and the current state of capitalism.

---

<sup>1</sup> ETH Zurich

# **Intersectionality and Digital Humanities in the Teaching of Architectural History: Diversity in the Dissemination of Knowledge**

**Marianna Charitonidou<sup>1</sup>**

Taking as its starting point the increasing importance of the role of digital curators within institutions holding architectural archives, the article aims to elaborate tools coming from intersectional theory and practice in order to produce an understanding of how women and black men are represented in teaching architectural history in an ensemble of emblematic schools of architecture. More specifically, the paper, through the elaboration of concepts and tools coming from the theory of intersectionality, examine how aspects concerning gender and race can be taken into account when establishing a curriculum of teaching architectural history. It is based on the hypothesis that visualisation strategies can show the evolution of the role of women and black people in architectural discourse. Drawing upon Kimberlé Crenshaw's work, and on the impact of the theory of intersectionality on digital humanities and digital labour studies, the project aims to shape a method of digital curation able to conjointly address issues of race, gender, class, ability, sexuality, or other categories of difference while interpreting the primary sources. Particular emphasis is placed on the fact that the intersectional perspective is the endeavour to interrogate its own positionality and the very processes of knowledge production, the project also explores how visualisation strategies can show the evolution of the role of women and black people in architectural discourse. A seminal text by Crenshaw, which is of great significance for the project, is her article entitled "Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color", published in *Stanford Law Review* in 1991. In this article, Crenshaw argued that "both women and people of color" are marginalized by "discourses that are shaped to respond to one [identity] or the other" (Crenshaw 1991), rather than both.

Most recently, the theory of intersectionality was introduced into the digital humanities in order to address issues regarding gender and race conjointly. As far as the field of architecture is concerned, the question of race is becoming more present in ongoing debates, as is evidenced by the recently published book *Race and Modern Architecture: A Critical History from the Enlightenment to the Present* (2020), edited by Irene Cheng, Charles L. Davis II and Mabel O. Wilson, and projects such as the *Black Architects Archive (BAA)* by Jay Cephas, whose aim was to collect and display the work of Black architects across history in an effort to bring to light underrepresented

---

<sup>1</sup> ETH Zurich

practitioners in architecture. The same is valid for the question of gender, as appears through the organisation of events including the symposium “The Fielding Architecture: Feminist Practices for a Decolonised Pedagogy”, which took place at the University of Brighton in June 2019, and the emergence of collectives such as Feminist Art and Architecture Collaborative, which in its manifest published in the Harvard Design Magazine describes itself as “a transnational coalition of feminists, awake to [...] [their] positioning as “Others” within the patriarchy; awake to [...] [their] exclusion from unmarked norm(s), awake to [their] [...] emergence from a history of subjugation, subordination, and colonization” (FAAC 2018).

Starting out from the hypothesis that it is becoming increasingly necessary to address these issues conjointly in the ongoing architectural debates, the paper presents certain methods of teaching architectural history that intend to bring the aforementioned aspects together. An important benefit of tackling gender and race issues simultaneously is the capacity to “address the structural parameters that are set up when a homogeneous group has been at the center and don’t automatically engender understanding across forms of difference”, as Moya Bailey has argued (Bailey 2020). Another noteworthy characteristic of the intersectional perspective is the endeavour to interrogate its own positionality and the very processes of knowledge production.

### Selective References

- Bailey, Moya, “All the Digital Humanists Are White, All the Nerds Are Men, But Some of Us Are Brave”, in Barbara Bordalejo, Roopika Risam, eds., *Intersectionality in Digital Humanities* (Amsterdam: Arc Humanities Press, 2020) 9-12.
- Bilge, Sirma, “Intersectionality undone: Saving intersectionality from feminist intersectionality studies”, *Du Bois Review*, 10(2) (2013): 405-424.
- Carastathis, Anna, *Intersectionality: Origins, Contestations, Horizons* (Nebraska: University of Nebraska Press, 2016).
- Cheng, Irene, Charles L. Davis II, Mabel O. Wilson, eds., *Race and Modern Architecture: A Critical History from the Enlightenment to the Present* (Pittsburgh: University of Pittsburgh Press 2020).
- Collins, Patricia Hill, Sirma Bilge, *Intersectionality* (Cambridge: Polity Press: 2016).
- Cooper, Brittney, “Intersectionality”, in Lisa Disch, Mary Hawkesworth, eds., *The Oxford Handbook of Feminist Theory* (New York: Oxford University Press, 2016).
- Crenshaw, Kimberlé, “Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color”, in *Stanford Law Review*, 43(6) (1991): 1241-1299.
- Doyle, Shelby, Leslie Forehand, “Fabricating Architecture: Digital Craft as Feminist Practice”, *the Avery Review*, 25 (2017): 1-10.

FAAC, "To Manifest", *Harvard Design Magazine* 46: No Sweat (2018): 182-189.

Harris, Jessica C., Lori D. Patton, "Un/Doing Intersectionality through Higher Education Research", *The Journal of Higher Education*, 90(3) (2019): 347-372.

Marie, Jakia, Donald "DJ" Mitchell Jr., Tiffany L. Steele, *Intersectionality & Higher Education: Research, Theory, & Praxis* (New York: Peter Lang, 2019).

Romero, Mary, *Introducing Intersectionality* (Cambridge: Polity Press, 2018).

# Multilingual word embeddings and low resources: identifying influence in Antiquity

Marianne Reboul, *École Normale Supérieure de Lyon*

## Background

The computing of semantic influence is a key part to understanding the development and movement of knowledge throughout the world. It is now made possible by the use of powerful computer hardware, and the correlated development of new machine learning techniques. Semantic influence therefore becomes a relationship to be quantified, graphically represented and analyzed: it is now possible to measure how meaning was altered within a language and between languages.

Although new techniques allow researchers to establish semantic proximity between languages using machine learning for aligning and predicting translations, those techniques are data hungry and tend to be less effective on very low resource languages such as ancient languages. Most of the techniques used to develop translation models rely on parallel data training: this is partially due to the fact that automatic translation is built for modern, period independent, usage, such as commercial translations. However, the lack of parallel data for rarer languages and ancient languages has led to research in using monolingual data training for multilingual vector spaces. Recent developments can overcome the scarcity of data using neural nets to produce linear mapping between languages. In this paper, we would like to apply techniques we tested using monolingual training and mapping to create low resource multilingual embeddings without using massive parallel data. We would both use semi-supervised and fully unsupervised techniques.

## Isomorphic semantic spaces

Our research relies on an instinctive, yet bold assumption that Ancient European languages are isomorphic, that is to say that we assume that we could show most languages were built in comparable semantic molds, and share, to a certain extent, part of the meaning of their words, or at least share a common semantic conception of their world. The models obtained while training on monolingual data should therefore be comparable. We would firstly explain how, based on the assumption that ancient languages are isomorphic, monolingual semantic spaces can be mapped in one single multilingual space. Secondly, we would demonstrate that semantic spaces trained on different time periods can reveal potential intertextuality and semantic influences between texts. Thirdly, we would apply those techniques to a specific object, that is to say a corpus of Greek and Latin epic poems,

from Homer to Vergil, and see how specific terms and aspects of Latin epics tend to be more influenced or disassociate themselves from the Greek canon.

### Methods and preliminary results

For our preliminary results, we used two kinds of corpora, although both could be qualified as microcorpora : the training corpus is a compilation of both Homer's *Iliad* and *Odyssey* (Allen's Greek version (Allen, 1908), Estienne's Latin version (Estienne, 1589)), and a test corpus of ten to fifteen texts (TEI format, already lemmatized by Perseus (Smith et al., 2000) or lemmatized in python through pie-extended (Manjavacas et al., 2019)), both in prose and verse, of canonical Greek and Latin epics, such as Vergil's Aeneid or Lucan's *Pharsalia* (although certain texts are not, strictly speaking, epic poems). We also built a very small custom Greek to Latin dictionary, based on two French dictionaries (although we intend to modify this step in future experiments), for the training phase.

If we consider Greek and Latin as isomorphic languages, we would expect some of the terms used in one language to find their most probable correspondance in the other according to their context. That is to say, their semantical environment should be approximately comparable. For example, "Jupiter" and "Ζεύς" should share space with comparable semantic terms used in the same context. If we try to describe this phenomenon with a 3-dimensional semantic space (although our spaces generally have 200 dimensions), one should then see semantic spaces in both languages as two clouds of terms having roughly the same aspect. The method to associate both spaces (in our case trained either with GloVe (Pennington et al., 2014) or Word2Vec (Mikolov et al., 2013)) would be to move one target space to fit the other. Several methods exist to do so, but the one that gave most effective results on our corpora is the VecMap (Artetxe et al., 2018) method, using linear transformation to map isomorphic spaces. Once the crosslingual semantic space has been created, we can see what vectors one language is more likely to share with the other. However, this is not yet an understanding of "influence", strictly speaking, as it lacks the aspect of evolution through time. This is the third part of the training (after monolingual training and crosslingual training), that is to say historical linking between crosslingual semantic spaces. Our results are not definite at this point yet, but we intend to use dynamic link prediction algorithms (eg. methods described in (Kutuzov et al., 2018)) to measure the evolution process of corsslingual vectors through time.

In the following example (Figure 1), which is a representation of random crosslingual vectors found on two microcorpora (Homer's Greek *Iliad* and Sommer's French *Odyssey*), one can see the accuracy of corsslingual associations on very scarce data. Further figures will be included during the presentation (which were not included here due to black and white publication guidelines).

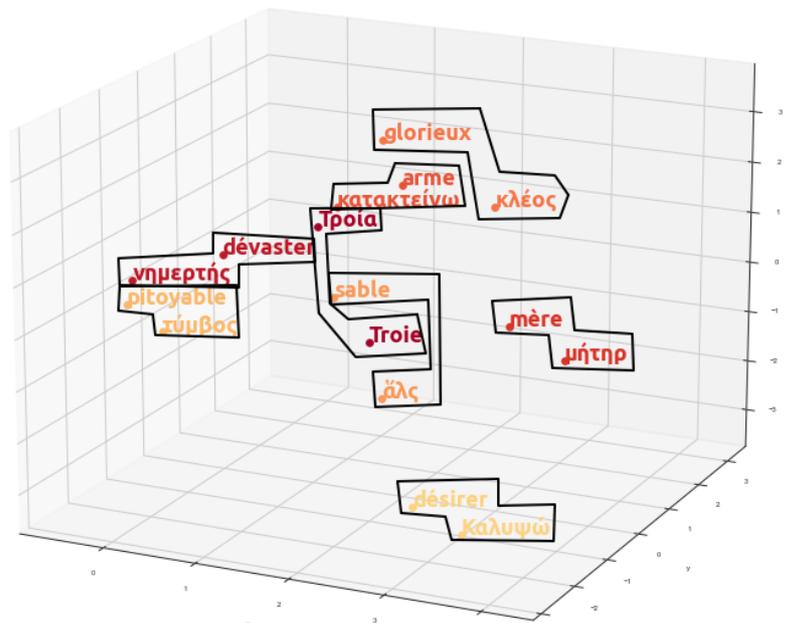


Figure 1 : Preliminary crosslingual semantic space (French-Greek) with nearest neighbours highlighted

### Future prospects

This paper would be a first step towards a larger project we intend to develop in the future, which could be summed up in three points : we would identify semantic influences in Ancient European written languages on one another first by building a corpus, in standardized TEI, of at least 50 million words (from 1450 B.C. to 400 A.C) first in Greek, Latin, then extending it to oldest and rarest European written languages (such as Etruscan, Minoan, Iberian, Linear B) ; we then would quantify semantic influences by computing a multilingual semantic space using crosslingual sentence and word embeddings from monolingual pre-training ; lastly, we would show these influences by analyzing and graphically representing chronological correlations between those languages. I therefore hypothesize that I can measure and show the interaction and influence of European Ancient written languages on each other through time.

### References

- [1]. ALLEN, Thomas William, *et al.* (ed.). *Homeri Ilias*. Clarendon Press, 1908.
- [2]. ALLEN, Thomas William, *et al.* (ed.). *Homeri Opera, Tomus III, Odysseae*. Clarendon Press, 1961.
- [3]. ESTIENNE, Henri, *Homeri Poemata duo, Ilias et Odyssea, sive Ulyssea*, Genève, Estienne, 1589.

- [4]. SMITH, David A., RYDBERG-COX, Jeffrey A., et CRANE, Gregory R. The Perseus Project: A digital library for the humanities. *Literary and Linguistic Computing*, 2000, vol. 15, no 1, p. 15-25.
- [5]. MANJAVACAS, Enrique, KÁDÁR, Ákos, et KESTEMONT, Mike. Improving lemmatization of non-standard languages with joint learning. *arXiv preprint arXiv:1903.06939*, 2019.
- [6]. PENNINGTON, Jeffrey, SOCHER, Richard, et MANNING, Christopher D. Glove: Global vectors for word representation. In : *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014. p. 1532-1543.
- [7]. MIKOLOV, Tomas, CHEN, Kai, CORRADO, Greg, et al. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- [8]. ARTETXE, Mikel, LABAKA, Gorka, et AGIRRE, Eneko. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *arXiv preprint arXiv:1805.06297*, 2018.
- [9]. KUTUZOV, Andrey, ØVRELID, Lilja, SZYMANSKI, Terrence, et al. Diachronic word embeddings and semantic shifts: a survey. *arXiv preprint arXiv:1806.03537*, 2018.

# Skin Deep: Exploring Ideals of Japanese Beauty through Social Media

Amy Grace Metcalfe<sup>1</sup>, Emily Ohman<sup>1</sup>

## Introduction

Instagram is one of the most widely used social media platforms in Japan, with an audience of over 38 million [13]. Instagram's focus on photo sharing generates a saturation of images conveying beauty. Since Naomi Wolf's 1991 publication "The Beauty Myth", countless studies have shown the negative effects the beauty industry has on women, including obsessions over weight [7], anxiety [9], and furthering inequality between the sexes [15]. This study examines the linguistic features of Instagram posts by beauty companies, both qualitatively and quantitatively, in order to discover how language contributes to the construction of beauty standards in the Japanese context. We use established natural language processing and topic modeling, combined with critical discourse analysis (CDA) to understand both the quantifiable data and the social effects of these Instagram posts. This study contributes to a better understanding of the definition of beauty within the Japanese context and offers additional insights into beauty ideals and how these are fabricated.

The objectivity of qualitative analyses have been criticised for being too subjective [3]. Whether this is entirely justified or not (see e.g. [2]), qualitative analyses supported by quantitative methods have proven useful in investigating media discourse, allowing researchers to identify textual [12]. Incorporating NLP and topic modeling provides scaffolding for the CDA, leading to verifiable empirical results.

## Data & Methods

The data was collected using the Instaloader package for Python. The Instagram profiles of twenty companies were randomly chosen, with the criteria being that the company's product or service must be available in Japan and that they have an Instagram profile aimed at the Japanese market. In the case where the brand has many profiles, the profile mainly written in Japanese was selected. The companies chosen were; Beauteen (beauteen\_offical), Chifure (chifure\_official), Curél (curel\_official\_jp), DHC (dhc\_official\_jp), Etude House (etudejapan), Ichikami (ichikami\_kracie) , Innisfree (innisfreejapan), Kanebo (kaneboofficial), Kireimo (kireimo\_official), Kosé (kose\_official), Liese (liese\_official\_jp), Maybelline (maybellinejp), Musée

---

<sup>1</sup> Waseda University

(museeplatinum\_insta), Palty (paty\_official), Revlon (revlonjapan), Rimmel (rimmellondon\_jp), Rize Clinic (rizeclinic), Sekkisei (sekkisei.official), Shiseido (shiseido\_japan), and TBC Aesthetic (tbc\_aesthetic). Our data consists of 7477 individual Instagram posts.

The json-formatted data was then converted to dataframes for exploratory data analysis. The posts themselves were segmented, and annotated for part of speech using Fugashi, mecab, and Spacy. This gave us access to both the simplified Universal Part-of-Speech tagset used by SpaCy such as ADJ - adjectives, but also Japanese tags which made it possible to include adjectives that can in context be nouns in addition to 形容詞 tags in our keywords, and therefore not miss important keywords. As our focus is words describing beauty, adjectives are an important part of our analysis. The raw text was then cleaned up and used for LDA-based topic modeling using *dariah*.

Primarily Fairclough's three-part model has been used in the theoretical framework (see e.g. [6,8]), with the micro, meso, and macro levels of interpretation allowing for the linguistic features of the text, the strategies used and how the message is conveyed, and social effects to be considered respectively [4].

## **Results, Analysis & Discussion**

From a preliminary analysis of the word frequency table, there are two trends; the frequency of the brand name, and the prevalence of words which can be grouped together into the category of skincare, these groupings are also evident in the topics that emerge through topic modeling the posts.

The brand name, both in roman letters and katakana, are among the most common words in our corpus, making up just under 25% of the first one hundred most frequent words, lending support to the notion that brands have their own aesthetic value and culture [14]. These beauty brands are attempting to participate in the mania surrounding branded goods by accentuating their name in social media posts.

Furthermore, much emphasis is placed on the care and maintenance of the skin (see table 1). It cannot be ignored that within the Japanese context, skin care and skin whitening practices tend to coincide [1]. Evidence from the corpus suggests that such ideals are still popular among consumers today, with whiteness explicitly mentioned with the words 雪肌精 and 美白.

tokens	keyword	translation
95348	美白	whitening
94349	skincare	skincare
93351	乾燥肌	dry skin
85369	ファンデーション	foundation
76398	保湿	moisturizing
66464	クリーム	cream
61849	スキンケア	skincare
56509	肌ケア	skincare
55518	雪肌精	"snow skin" essence
48585	美肌	beautiful skin
36699	skin	skin

Table 1: Summary overview of common tokens in corpus

Overall, patterns that emerge from the corpus, reveal elements of what beauty means in the Japanese context, namely the inclusion of brand equity, and the importance of the skin, as well as white skin still as the ideal (see e.g. [5]). Different companies use adjectives differently and these groupings can be seen in both the list of adjectives and adverbs used to describe these ideals as well as in the topics that emerge from topic modeling. Some of the more interesting topics include COVID-related issues such as mask-specific makeup (i.e. マスクメイク、新商品、美肌作り) and how to take care of yourself and your skin (丁寧な暮らし、暮らしを楽しむ、贅沢な時間、美容好きな人と繋がりたい) during these times. The push and pull of the internet can lead to a cycle where companies set the narrative that consumers are more than happy to partake in regardless of their own ideals (cf. [11]). We expect this corpus to yield many more interesting insights into Japanese beauty ideals, COVID-related self-care, and societal issues involving the pressure to conform to Japan-specific beauty standards.

## References

- [1]. Ashikari, M., (2005) "Cultivating Japanese Whiteness: The 'Whitening' Cosmetics Boom and the Japanese Identity.", *Journal of Material Culture*, 10(1), pp. 73-91.
- [2]. Baškarada, S. and Koronios, A., (2018) 'A philosophical discussion of qualitative, quantitative, and mixed methods research in social science.', *Qualitative Research Journal*, 18(1),

- [3]. Cheng, W., (2013) 'Corpus-Based Linguistic Approaches to Critical Discourse Analysis', *The Encyclopedia of Applied Linguistics*, pp. 1-8
- [4]. Fairclough, N., (2001) *Language and Power*, New York: Longman Inc.
- [5]. Grinschpun, H., (2012) 'The City and the Chain: Conceptualizing Globalization and Consumption in Japan', *Japan Review*, 24, pp. 169-195
- [6]. Kaur, K., Arumugam, N., and Yunus, N.M., (2013) 'Beauty Product Advertisements: A Critical Discourse Analysis', *Asian Social Science*, 9(3), pp. 61-71.
- [7]. Kayano, M., et al., (2008) 'Eating attitudes and body dissatisfaction in adolescents: Cross-cultural study', *Psychiatry and Clinical Neurosciences*, 62, pp. 17-25.
- [8]. Lestari, E.M. I., (2020) 'A Critical Discourse Analysis of The Advertisement of Japanese Beauty Products', *IZUMI*, 9(1), pp. 2502-3535.
- [9]. Miller, L., (2006), *Beauty Up: Exploring contemporary Japanese body aesthetics*, Los Angeles: University of California Press.
- [10]. Miller, L. (2004) 'Youth fashion and changing beautification practices', Matthews, G., White, B. (eds.). *Japan's Changing Generations; Are young people creating a new society?*, Abingdon: Routledge, pp. 83-97.
- [11]. Nathan-Tilloy, C., Shann, G., Skea, B., (2016) *The Dove Global Beauty And Confidence Report*, Dove.
- [12]. O' Halloran, K., (2010) 'How to use corpus linguistics in the study of media discourse', in O' Keefe, A., and McCarthy, M. (eds.), *The Routledge Handbook of Corpus Linguistics*, London and New York: Routledge
- [13]. "Leading countries based on Instagram audience size as of January 2021." *Statistica*, January 2021, <https://www.statista.com/statistics/578364/countries-withmost-instagram-users/> (Accessed: 5 June 2021)
- [14]. Toffoletti, K., and Thorpe, H., (2018) 'The athletic labour of femininity: The branding and consumption of global celebrity sportswomen on Instagram', *Journal of Consumer Culture*, 18(2), pp. 298-316
- [15]. Walter, N., (2010) *Living Dolls: The Return of Sexism*, London: Virago.

# Analyzing “Mechanisms” in the British National Corpus

Yuki Sugawara<sup>1</sup>

## Introduction

The words "mechanisms" are pervasive not only in academic contexts but also in daily life. In addition, some philosophers of science argue that the concept of "mechanisms" has a vital role in understanding scientific practices (Machamer et al., 2000; Glennan 2002). This paper will analyze how the words "mechanisms" are used in the British National Corpus (BNC) that includes large-scale language practices in various genres. This paper will pick up all sentences that include the verbs with the words of "mechanisms" as the object, pick up the verbs, and describe their characteristics in the context of both genres and actual sentences. In based on them, This paper will categorize their types of verbs as (a) descriptive-type, (b) investigative-type, (c) relative-type, and (d) discover-type, and show that (a) descriptive-type is the most central and the others are peripheral.

## Methods

This paper used the corpus search application called "Sketch Engine" to pick up data from the BNC. The Sketch Engine provides 487 corpus worldwide and various advanced searches (Kilgarriff et al. 2004; Kilgarriff et al. 2014). This paper picked up all cases that include the verbs with the words "mechanisms" as the object from the BNC, using the Concordance function. Although the words "mechanisms" are used 4921 times in the BNC, their words are used 1270 times as the object of the verbs. This paper picked up 356 cases that are ranked in the top 25 as frequency. In analyzing their distribution, This paper used the behavioural profile investigated as a method in corpus linguistics (Gries and Otani 2010). The behavioural profile includes making a dataset for describing various language practices to capture the behaviour of words. This paper did a cluster analysis on its dataset using R Core Team 2019 and used the Canberra to calculate the distance between items and the Ward as between clusters. After clustering, this paper did a qualitative analysis of actual texts.

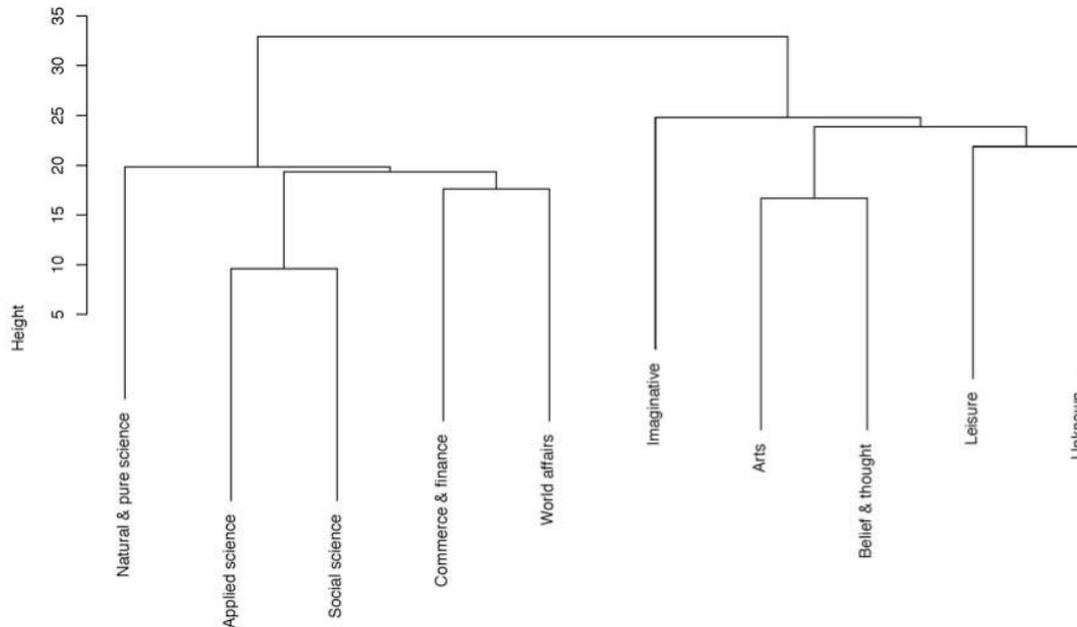
## Results

The results of cluster analysis are shown in Figure 1. This paper calls the left side of the cluster "the descriptive cluster" and the right side of the cluster "the sub-descriptive cluster". The differences between them are (i) the frequency of the verb "provide", (ii) the frequencies of the verbs that are ranked in the top 25, (iii) the number of the verbs that the

---

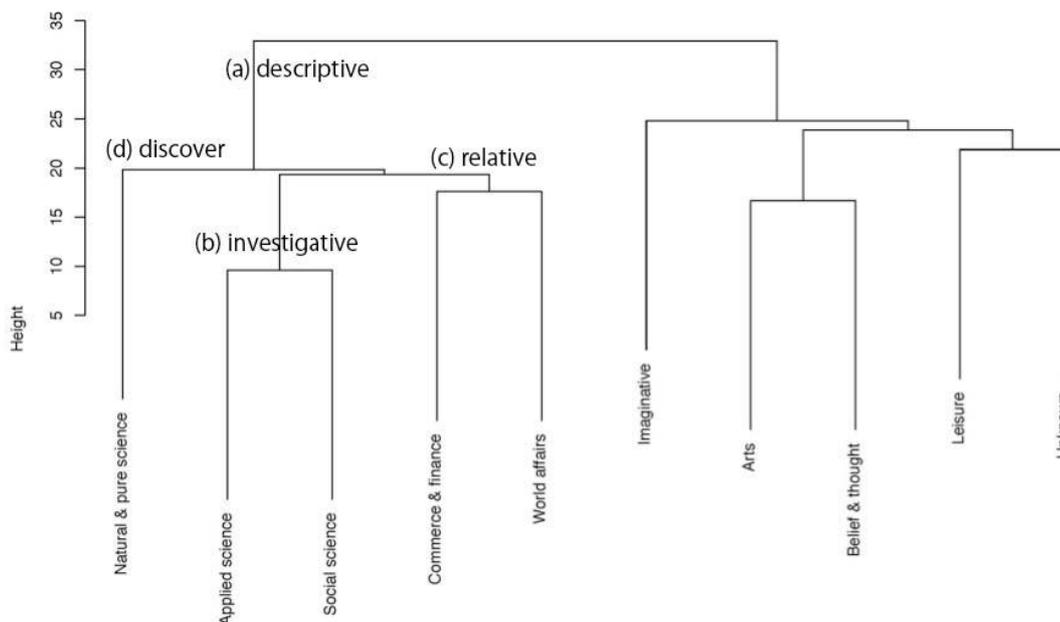
<sup>1</sup> Keio University

relative frequencies exceed 0.3. The similarities between them are (i) the verb of "provide" co-occur in all genres except for the genre of Leisure, Unknown, and Imaginative (I will characterize these behaviours as (a) descriptive-type). (ii) the verbs such as "understand" and "find" co-occur in the peripheral clusters ("Natural & pure science" and "Imaginative") that combine lately.



**Figure 1: Cluster dendrogram.**

The characteristics of these clusters are shown in Figure 2. The types in Figure 2 are characterized by these verbs that are used in the clusters (Table 1). For example, in the genres of "Applied Science" and "Social Science", the verbs such as "investigate" and "require" co-occur. In the genres of "Commerce & finance" and "World affairs", the verb "join" co-occur. In the genre of "Natural & pure science", the verbs such as "find" and "understand" co-occur. Based on these observations, this paper will characterize the clusters of "Applied Science" and "Social Science" as (b) investigative-type, the clusters of "Commerce & finance" and "World affairs" as (c) relative-type, and the cluster of "Natural & pure science" as (d) discover-type.



**Figure 2: Coded cluster dendrogram.**

**Table 1: The types of Verbs.**

(a) descriptive	(b) investigative	(c) relative	(d) discover
provide	investigate	join	find
	require	use	understand
	develop	establish	reveal
	elucidate	enter	
	devise	introduce	
	reveal	create	
	examine	contain	
	study	examine	
	suggest		
	need		
	see		

**Reference**

[1]. Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of science*, 67(1), 1-25.

[2]. Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of science*, 69(S3), S342-S353.

[3]. Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). Itri-04-08 the sketch engine. *Information Technology*, 105(116).

[4]. Kilgarriff, A., Baisa, V., Bušta, J., Jakubiček, M., Kovář, V., Michelfeit, J., ... & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), 7-36.

[5]. Gries, S. T., & Otani, N. (2010). Behavioral profiles: A corpus-based perspective on synonymy and antonymy. *ICAME Journal*, 34, 121-150.

## **One Challenge, Not Two Problems: Regular Expressions for Researching a Single-Author Corpus**

**Dr. Robert W. Williams<sup>1</sup>**

My project aligns with an interpretive approach to the study of ideas: I seek to comprehend how humans express their understandings of the world in the meaningful words they write. I am researching the concept of the “unknowable” in the thought of W.E.B. Du Bois (1868-1963), an African American civil rights activist and scholar. I created a 230-document corpus containing over 3 million words encompassing his essays, newspaper articles, drafts, and 19 books. This corpus is neither comprehensive nor fully representative of his over 2000 published writings.

In the world of Du Bois scholarship his idea of “unknowable” is understudied, yet as I have argued in my scholarly work ([12]), it is a significant theme in his own research and activism. For Du Bois, the unknowable was not merely a lack of personal or general knowledge about something. Rather, it involved a profound lack of knowledge because data was (and is) epistemologically inaccessible. This inaccessibility is due, for example, to our inability to know about events because of irretrievable information, and to our inability to know others directly because we cannot experience what they are feeling.

My project's specific goal is to find Du Bois's concept of the “unknowable” as well as its variant expressions amid the myriad words of the corpus. From any such matches, the second goal is to discern how Du Bois defines and applies the concept.

This interpretive project requires two vital elements. First, I seek the synonymous words by which Du Bois expressed the concept of the “unknowable.” I must be able to access in the corpus the richness of Du Bois's nuanced, aesthetic, and theoretical expressions. Second, I need sufficient textual details in order to disambiguate the uses of words and phrases that any computational technique presents as output. I prefer to disambiguate by reading more closely into the word's context—the sentences, paragraphs, and even the document itself. This accords with my vocation as a political theorist: to minutely examine the intricacies of Du Bois's multiple expressions of the “unknowable,” even the unique ones, the hapax legomena. Digital humanities techniques that reduce the lexicon to summaries or to statistically derived measures—as fruitful as they are for other projects—do not suffice for mine.

Accordingly, a concordancer (here AntConc) is my software tool of choice for researching a corpus. In order to locate the words salient to interpreting the “unknowable” concept, my technique of choice involves regular expressions (regexes), a

---

<sup>1</sup> Formerly: Bennett College (USA); [drrobtwms3@hotmail.com](mailto:drrobtwms3@hotmail.com); [www.webdubois.org](http://www.webdubois.org)

notational system permitting us to match patterns of characters, such as a word or even words near each other, within larger spans of text.

### Workflow

My presentation sketches the reiterative workflow of my interpretive process, which includes these steps:

**A.** Search for the node word “unknowable” or “un-knowable”, and “unknown”:

(?i)un-?knowab

(?i)¥bunknown

“Un(-)knowable” is scarce in the corpus ([5: Ch.X], [6], [7], [8], [9], [10]). “Unknown” as unknowable is somewhat more prevalent (e.g., [1], [3]).

**B.** Within the co-text presented in the concordance lines and the etexts, by which words did Du Bois define, apply, or adapt “unknowable”? For example, the following co-occurring words offer us potentially fruitful node words with which to search ([5: Ch.VI], [6], [8]):

limitations limit

science scientific sciences scientifically scientist

knowledge

reasonable reason reasoning

logic logical

fiction

history historical

facts

imagination

own only

**C.** Craft regexes to search for node words derived from the co-text. For example,

(?i)¥bknow

We can even search for words in the co-text that are not variants of “know”, but which relate to it, such as “science” or “scientific”, “reason”, “logic”, and so forth. For example:

(?i)¥bscien

**D.** When reading through the concordance lines and the full-text for the node words, we also locate multi-word phrases that potentially convey or apply the idea, including

“will/can never know”

“can never be known”

“none will ever know”

“These facts are gone forever.” ([8])

“only the man himself. . .knows his own condition” ([5: Ch.VI])

For multi-word phrases we can use proximity-oriented regexes to locate nearby words within the documents, including a reverse order of the words. For example:

```
(?i)(not|never)(?:.){1,30}?\bknow
(?i)\bknow(?:.){1,30}?(not|never)
(?i)know(?:.){1,30}?(only|alone|own)
(?i)(only|alone|own)(?:.){1,30}?know
```

## Discussion

**First:** When Du Bois used the specific term “unknowable” in the few places it did appear it was often related to knowledge in general or to historical research in general ([6], [7], [8: “Postscript”]). Only in a relatively smaller number of cases did Du Bois use “unknown” to refer to unknowability (e.g., [1: missing population data *passim*]).

**Second:** When Du Bois utilized specific multi-word phrases to express unknowability, such as “will/can never know” or “none will ever know,” Du Bois was focusing on particular instances where specific pieces of information were unrecoverable (e.g. [2: ¶27]). On the other hand, in most cases “do not know” referred to few instances of the unknowable in principle, with a notable exceptions being “The Individual and Social Conscience” ([4: ¶3]).

**Third:** When Du Bois was indicating that the unknowable involved no direct knowledge of another's thoughts, experiences, and feelings, then words like “alone”, “only”, and “own” occurred in the near vicinity of the lemma “know” (e.g., [5: Ch.VI]).

**Fourth:** For Du Bois writing the past tense of “knew” as part of a phrase “knew not” and “knew nothing” did not invoke unknowability in principle. Persons did come to know or could have known in and through other circumstances. One exception was the adverb “never” associated with “knew”. Here “never knew” did tend to implicate unknowability in principle ([11: p.120]). Reading the concordance lines helped to disambiguate the cases.

Ultimately, such concordancer-mediated techniques navigate between distance and closer forms of reading. They help us to study how authors articulate their ideas in multifarious ways within the individual texts of a corpus.

## References

- [1] Du Bois, W.E.B. 1899. *The Philadelphia Negro*. Philadelphia: Ginn.
- [2] ----- . 1904. “The Development of a People,” *International Journal of Ethics*, 14:3 (April): 292-311.
- [3] ----- . 1904. “Heredity and the Public Schools.” Pp.45-52 in *Pamphlets and Leaflets*. Herbert Aptheker (Ed.). White Plains, NY: Kraus-Thomson Organization, 1986.

- [4] -----, 1905. "The Individual and Social Conscience" [Originally Untitled]. Pp.53-55 in Religious Education Association, *The Aims of Religious Education. The Proceedings of the Third Annual Convention...*, 1905. Chicago: Executive Office of the R.E.A.
- [5] -----, 1920. *Darkwater*. NY: Harcourt, Brace and Howe.
- [6] -----, 1943. "Letter from W.E.B. Du Bois to American Philosophical Association, December 13, 1943." W.E.B. Du Bois Papers. Special Collections & University Archives. University of Massachusetts Library. <<http://credo.library.umass.edu/view/full/mums312-b099-i286>>.
- [7] -----, 1956. "Letter to Herbert Aptheker, January 10, 1956." Pp.394-396 in *The Correspondence of W.E.B. Du Bois: Vol. III: Selections, 1944-1963*. Herbert Aptheker (Ed.). Amherst: University of Massachusetts Press, 1978.
- [8] -----, 1957. *The Ordeal of Mansart*. NY: Mainstream Publishers.
- [9] -----, 1959. *Mansart Builds a School*. NY: Mainstream Publishers.
- [10] -----, 1961. *Worlds of Color*. NY: Mainstream Publishers.
- [11] -----, 1968. *The Autobiography of W. E. B. Du Bois*. NY: International Publishers.
- [12] Williams, Robert W. 2018. "A Democracy of Differences: Knowledge and the Unknowable in Du Bois's Theory of Democratic Governance." Pp.181-203 in Nick Bromell (Ed.), *A Political Companion to W.E.B. Du Bois*. Lexington: University Press of Kentucky.

# Picking out Arabian Names from *Fahrassa* by Ja‘far b. Idrīs al-Kattānī without Reading Arabic

Yuri Ishida<sup>1</sup>, Kensuke Baba<sup>1</sup>

## Introduction

What we call “radical” Islam is characterized by a tendency to literally interpret the Quran. Its direct origin could be traced back to the Islamic reformism from the eighteenth century. Although the center of this movement was Mecca and Medina in the Arabian Peninsula, reformists from Morocco to Indonesia were united by the master–disciple relationship called “intellectual family tree” or “academic community” (Voll, 1975). To analyze this global network, a database of Islamic scholars is necessary. Building databases requires human resources. However, in Japan, there are only a few people who can read Arabic text. Besides outsourcing from Arabic nations, is there another way to build a database?

## Text

As a first step in automatically detecting scholars’ personal information (name, birth year, birthplace, etc.) from Arabic texts, we performed the machine learning on Arabian names in *Fahrassa* by Ja‘far b. Idrīs al-Kattānī (1830/1831–1905). He belonged to a celebrated family in Fez, a city in today’s Kingdom of Morocco, and his academic chains were connected to the Islamic reformists. As published by his descendants who praise his intellectual legitimacy, *Fahrassa* gives us his academic background, his masters, his disciples, books from which he learned, and so on. The meaning of *fahrassa* in Arabic is index, catalog, or list, and this title thus describes the book’s characteristics. The Arabic text is not very complicated, and the names of Islamic scholars are listed in the sections by field of study.

## Approach to find Arabian Names

A well-known joke is that if you call out “Muḥammad” on the street, several men will answer. In fact, Muḥammad was the most popular name for male newborns in the UK in 2014 (Baby Center, n.d.). Muslim names do not have many varieties, and they follow several rules. For example, some Islamic scholars indicate their hometown at the end of their name: al + the place name + ī. Thus, a man who came from Baghdad is named al-Baghdādī. However, names of places are infinite, and to find such names in the text, Arabic literacy is necessary.

---

<sup>1</sup> Okayama University

In *Fahrassa, al-Iḥyā'* of al-Ghazālī (1058–1111), one of the most eminent Islamic books, was transmitted via an intellectual chain from the author to al-Kattānī and involving 15 other scholars. The chain is connected by the word “from” (عن, *an*): “I [al-Kattānī] quote the book *al-Iḥyā'* written by al-Ghazālī and his other works from al-Shaykh ‘Alī b. Zāhir from al-Shaykh Aḥmad Minna Allāh al-Mālikī from.....al-Ḥāfiẓ Abī al-Faraj al-Baghdādī from the author [al-Ghazālī] of them [*al-Iḥyā'* and other books written by al-Ghazālī]” (al-Kattānī, 2004). Thus, in *Fahrassa*, “from” signals that the name of a scholar follows. Even if one does not read Arabic, one can recognize that “al-Shaykh ‘Alī b. Zāhir” is a person’s name because it is located between instances of “from.”

“From” appears approximately 1,500 times in the 46,500 words in *Fahrassa*, and there are other usages besides mentioning names. To specify the context of name enumeration, common names and honorary titles for Islamic scholars are also used. As mentioned above, Muḥammad is a popular name for Muslims, and there are many scholars named Muḥammad. In *Fahrassa*, Muḥammad, Aḥmad, ‘Alī, and Ḥasan are mentioned approximately 900, 300, 200, and 150 times, respectively. “Son” (*bun* or *ibn*) and “father” (*abī* or *abū*) are also strong candidates for names because of naming habits among Muslims. The former appears in *Fahrassa* approximately 2,000 times and the latter 800 times. Additionally, the honorific titles used for Islamic scholars are good supplementary identification information: “master” (*shaykh*) appears approximately 350 times, “religion” (*al-Dīn*) 200 times, and “leader” (*imām*) 100 times.

## Conclusion

In the part of *al-Iḥyā'* of al-Ghazālī, names of persons appear 18 times (the names of two people, al-Ghazālī and al-Anṣārī, are mentioned twice). The index at the end of the book only shows four people. If one reads Arabic characters, one might think there are 17 names because the form of “al + ... + ī” appears 17 times. However, there are five mistakes to be committed here: three people are missed and the same person is counted as two-person twice. To find a personal name depending on the formula “al + ..... + ī” does not work well because the names of places are unlimited, and there are some people whose name does not show the hometown.

Thus, this way calls for an ability to read Arabic characters. How about finding the word “from” (*an*) then? One finds 16 names because “from” appears 16 times. However, there are four mistakes: one finds neither al-Ghazālī twice nor al-Anṣārī once, and one might think “the author of them [the books]” (*mu'allif-hā*) is a personal name. In case one regards the term between “from,” one finds 15 names, committing the mistake three times.

To improve this method, unique terms for personal names can be introduced. In this part, “father” (*abī*) appears twice, “son” (*bun*) four times, and “master” (*shaykh*) twice. If these terms are used as additional layers, one can find al-Ghazālī this time and be confident in the names of six persons, and the same mistake happen in only two places. This presentation shows the best combination of these words based on the analysis for other parts of *Fahrasa*.

## References

**Baby Centre.** (n.d.). Top Baby Boy Names 2014.

<https://www.babycentre.co.uk/a25011625/top-baby-boy-names-2014> (accessed 15 August 2021).

**al-Kattānī, J. b. I.** (2004). *Fahrasa Ja‘far ibn Idrīs al-Kattānī*. Beirut: Dār Ibn Ḥazm, p. 227.

**Voll, J.** (1975). Muḥammad Ḥayyā al-Sindī and Muḥammad ibn ‘Abd al-Wahhāb: An Analysis of an Intellectual Group in Eighteenth-Century Madīna. *Bulletin of the School of Oriental and African Studies*, 38 (1): 32–39.

# POS tagging for Vedic Sanskrit using deep learning

Yuzuki Tsukagoshi<sup>1</sup>

## Introduction

In this study, we show a new approach to add part-of-speech (POS) tags to Sanskrit words in the Vedas. Many approaches to POS tagging for Sanskrit have been proposed up to now. However, all of them focus on Classical Sanskrit, and Vedic Sanskrit, the older layer of the Sanskrit language, was not considered by such studies. Therefore, to analyze Vedic Sanskrit, we should take account of the phonological, morphological, and syntactic differences from Classical Sanskrit. The crucial differences are the word accent and more complex verb conjugation. Therefore, the existing POS taggers cannot always analyze the sentences of the Vedas correctly. Furthermore, a rule-based approach for Vedic Sanskrit is not practical because studies show that neural network models perform far better than the rule-based (Srivatsava et al. 2018).

## Method

In this project, we create a neural network model for a POS tagging task by making use of HuggingFace Transformers, using the text of the Rigveda. The model is based on a pre-trained model and our formatted datasets, which are used for training and testing. With the help of the pre-trained model and the use of the state-of-the-art Transformer model, we can create a new model to assign POS tags to Vedic Sanskrit words relatively easily.

The Rigveda is the oldest Vedic literature, and a few versions of its electronic text is available. We take a version of VedaWeb (Kölligan et al. 2020) to prepare a training text that contains word forms, lemmata, and POS tags. In our experiment, the whole text of the Rigveda is employed because we already have a text with POS tags. Since few Vedic texts have POS tags as in Classical Sanskrit literature, the Rigveda is the only text on which we perform supervised machine learning.

We make a small change to the VedaWeb version of the Rigveda in order to add the necessary details. The VedaWeb text can give us morphological information which tells whether the word is noun, pronoun, verb, or indeclinable. However, the category of the word in the VedaWeb version is too broad, that is, there are only four sorts, nominal stem (= nouns and adjectives), pronoun, root (= verbs and verbal nouns), and invariable (or indeclinable) although there are many kinds of the information on nouns such as the case, the gender, and the number and on verbs about the tense, the mood, the voice, etc. There are three problems with this text. The first is that demonstrative, interrogative, relative, and personal pronouns are all put into the pronoun

---

<sup>1</sup> The University of Tokyo

category. In some cases, it cannot tell us whether a pronoun is, but we need POS tags showing this distinction. The second is that finite verbs and participles are put together in the verb category. However, we need to categorize them separately because participles in Sanskrit are used as nouns or adjectives, which have the case, the gender, and the number. The third is that adverbial nominal stems, adverbs, and negatives are all put in the indeclinable category. In such a text, we cannot even find out where the negative is placed in the sentence. In order to fill such gaps, we added more detailed information to the word category by assigning more appropriate labels to pronouns, verbs, and indeclinables and created a new POS tagged text of the Rigveda.

## Result

We trained the new model with HuggingFace Transformer and our arranged text, which took several days, and evaluated the results of the POS tagging of the test text acquired by our models. As a result, the precision is about 0.9704, the recall is about 0.9705, and the F1 score, which is the harmonic mean of the precision and recall, is about 0.9705.

## Conclusion and future work

Our new model produced a result that is accurate enough to be applied to other Vedic literature. Nevertheless, the language in the Rigveda is the most ancient language of (Vedic) Sanskrit and some studies distinguish it from the language of other Vedic literature.

In Sanskrit, there is a phenomenon called sandhi, in which the sound at the end of a word and/or at the beginning changes, and it is necessary to convert the changed word form that appears in the text into the word form before the change. Therefore, when we apply the present model to other Vedic texts, we also need to develop a procedure to undo the sandhi.

## Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 20J23373.

## Reference

- [1]. Kölligan, Daniel & Reinöhl, Uta (eds) in collaboration with Jakob Halfmann, Borge Kiss, Natalie Korobzow, Francisco Mondaca, Claes Neufeind & Patrick Sahle, with material provided by Paul Widmer et al. 2020. "VedaWeb - Online Research Platform for Old Indic Texts". Universität zu Köln. <https://vedaweb.uni-koeln.de>
- [2]. Srivastava, Prakhar, Kushal Chauhan, Deepanshu Aggarwal, Anupam Shukla, Joydip Dhar, and Vrashabh Prasad Jain. 2018. Deep Learning Based Unsupervised POS Tagging for Sanskrit. In *Proceedings of the 2018 International Conference on*

*Algorithms, Computing and Artificial Intelligence*. Association for Computing Machinery, New York, NY, USA, Article 56, 1–6.  
DOI:<https://doi.org/10.1145/3302425.3302487>

# Spectral analysis for identifying octave playing in piano works

Mai Takahashi<sup>1</sup>, Michikazu Kobayashi<sup>2</sup>, Ikki Ohmukai<sup>3</sup>

## Introduction

In this paper, we study interpretive editions in which authors have added their opinions to how to perform the work, and whether and how performers follow the “tradition” written in the interpretive editions by analyzing recorded data.

In conventional musicology, people had been more interested in critical editions which were aimed to be close to the composer's original intentions than the interpretive editions containing opinions from others [1]. However, new musicology starting around US and UK since 1990s have clarified that musical performances are affected not only by scores but also inheritances from masters to disciples [2,3], and people have begun to pay attention to interpretive editions which contains playing styles inherited continuously. On the other hand, it has been hardly clarified whether such playing styles in interpretive editions are truly practiced in performances.

In this research, we try to verify this by analyzing recorded data.

## Target in analysis

We focus on BWV903 Chromatic Fantasy and Fugue by J. S. Bach. There are nine interpretive editions published in between 1820 and 1928, the authors of which are related to each other as masters and disciples. Furthermore, C. Czerny, F. Liszt, and F. Busoni were ones of them: performers and educators who created playing styles being handed down to nowadays. Related to the editions, there are many recording data after 1912 by pupils of C. Czerny, F. Liszt, and F. Busoni.

The octave playing is one of traditions which are not appeared in urtext and critical editions of Bach but has been inherited through interpretive editions, i.e., a technique superimposing a lower octave on the bass voice. By analyzing 30 recorded data with pianos in between 1912 and 2015, we try to verify how the octave playing has been inherited. Although two of them were played by not pianos but piano rolls, their spectra of sounds are as sharp as those of pianos and we include them in our analyses. Because harpsichords and clavichords have broader spectra of sounds making analyses more difficult, we exclude recorded data with them from our analyses.

---

<sup>1</sup> The University of Tokyo

<sup>2</sup> Kochi University of Technology

<sup>3</sup> The University of Tokyo

## Methods of analysis

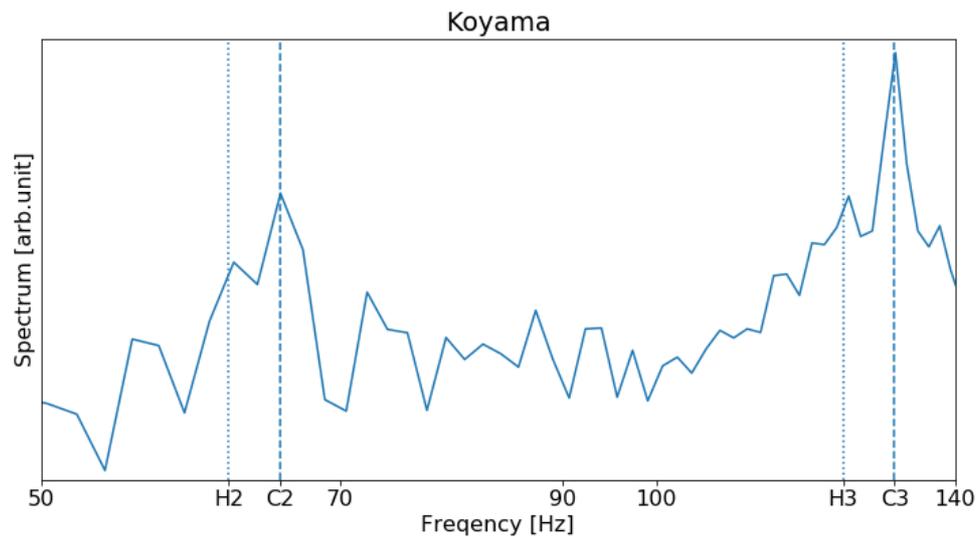
We analyze the first beat in bar 141 (the 421st beat) of Fugue (latter half of Chromatic Fantasy and Fugue), in which the lower octave playing in the bass voice is specified in the interpretive editions. This beat has a C3 (130.81Hz) sound in the urtext edition, and one octave lower C2 (65.4Hz) sound in the interpretive editions with the octave playing for the left-hand part. In order to accurately determine the beginning and the end of the 421st beat in the recorded data, we use Sonic Visualiser which enables to count beats for recorded data and export the beaten data as a text file [4]. Then, based on the beaten data, we extract the 421st beat from the recorded data and transform it into spectral data as a function of different frequencies via the Fourier transformation. Confirming a peak structure at frequencies around 123.47Hz and 65.41Hz, we attempt to determine whether the octave playing is practiced or not.

## Results

The octave playing is definitely confirmed in 12 recordings; Busoni, Giesecking, Kempff, Oborin, Weissenberg, Brendel, Sonoda, Koyama, Woodward, Delaage, Mejoueva, and Iwamoto. Figure 1 shows a characteristic spectrum for Koyama's playing. A clear peak structure can be confirmed not only at C3 which is indicated in the urtext edition, but also at C2, which is one octave below C3, clearly showing that this sound is also struck in the performance. In addition to C3 and C2 sounds, the peak structure can be seen at H3 and H2. This probably come from the previous 420th beat which contains H3 (and H2), and this sound remains due to reverberation.

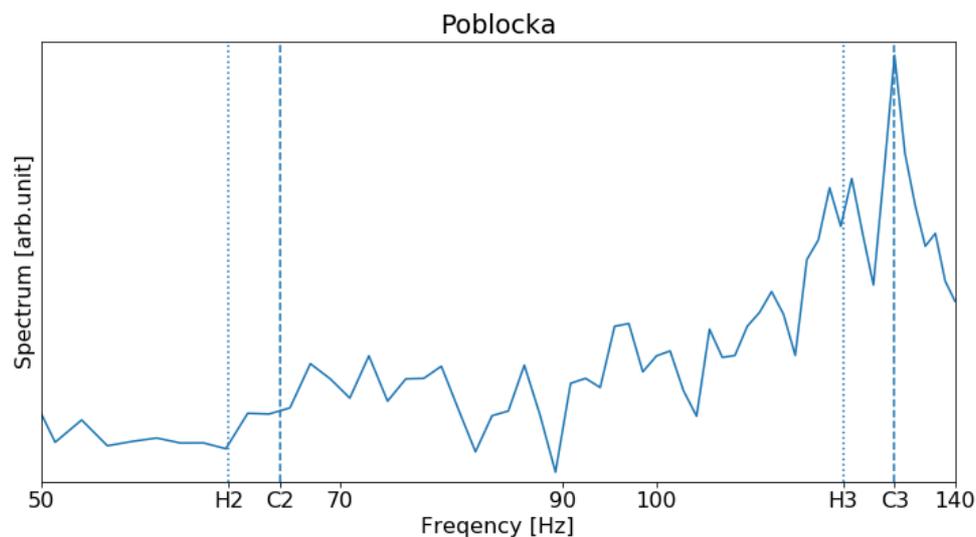
On the other hand, we can definitely confirm that the octave playing has not been practiced in 7 recordings; Arrau, Schnabel, Lifschitz, Poblocka, Pratt, Padova, and Kolly. Figure 2 shows a characteristic spectrum for Poblocka's playing. Different from Koyama's one shown in Fig. 1, there are only peaks at C3 and H3 and no peaks at C2 and H2.

For remaining 11 recordings, we cannot determine whether the octave playing is practiced or not in the current stage because peak structures at low frequencies are indistinct.



**Figure 1: Fourier spectrum of Koyama's performance from the beginning to the end of the 421st beat. Both horizontal and vertical axes are shown in logarithmic scales.**

**C3 (130.81Hz) and C2(65.41) sounds are indicated by dashed lines, and H3 (123.47Hz) and H2 (61.74Hz) sounds are indicated by dotted lines.**



**Figure 2: Fourier spectrum of Poblocka's performance from the beginning to the end of the 421st beat.**

It has been considered that publishing campaigns and spreading of the concept of urtext editions by musicologists [5] drove out interpretive editions rapidly [1]. However, our present work can reveal that the octave playing still survives toward today after 1940s.

We show all of our results in Table 1. Among 12 recordings with the octave playing, 7 recordings (Kempff, Weissenberg, Brendel, Oborin, Sonoda, and Koyama) are related with Czerny, Liszt and Busoni who have positively adopted the octave playing. Among 7 recordings without the octave playing, 3 recordings (Arrau, Poblocka, and Schnabel) are related to Liszt, Busoni, and Czerny. Although 11 recordings remain not to

be determined whether the octave playing is practiced or not, and we cannot give the clear conclusion at the current stage, there might be some relationship between Liszt, Busoni, and Czerny's groups and the octave playing.

**Table 1: Situation of octave playing**

year	performer	octave playing	year	performer	octave playing
1912	F. Busoni	Yes	1982	A. Schiff	Unclear
1912	W. Backhaus	Unclear	1987	I. Moravec	Unclear
1931	E. Fischer	Unclear	1988	K. Lifschitz	No
1945	C. Arrau	No	1991	E. Poblocka	No
1948	A. Schnabel	No	1995	A. Pratt	No
1948	W. Giesecking	Yes	1996	A. Padova	No
1948	E. Gilels	Unclear	1998	T. Sonoda	Yes
1950	M. Yudina	Unclear	1999	M. Koyama	Yes
1951	E. Petri	Unclear	2005	L. Fleisher	Unclear
1953	W. Kempff	Yes	2007	R. Woodward	Yes
1962	P. B.-Skoda	Unclear	2009	F. Delaage	Yes
1963	L. Oborin	Yes	2012	K.-A. Kolly	No
1966	A. Weissenberg	Yes	2013	I. Mejoueva	Yes
1973	A. Brendel	Yes	2014	M. Stadtfeld	Unclear
1982	T. Nikolayeva	Unclear	2015	W. Iwamoto	Yes

## Conclusion

By using the technique of informatics, we can objectively analyze the sound spectrum and clearly see that the performance style absent for urtext and critical editions has been inherited. Combining the analysis of interpretive editions, we can analyse 200 years of the performance styles, and analyses of recorded data are quite important in performance research.

## Reference

- [1]. H. Watanabe, Western Music: Preface to Performance History - Performance History of Beethoven Piano Sonata (Japanese), Shunju (2001), p. 160-161.
- [2]. N. Cook, Beyond the Score: Music as Performance, Oxford University Press, Oxford (2014), p. 163.
- [3]. H. Watanabe, Encyclopedia of Aesthetics (Japanese), Maruzen (2020), p. 396-397.
- [4]. <https://sonicvisualiser.org/> (2021/6/7 accessed).

- [5]. S. Ohsaki, Foration of Music History and Media (Japanese), Heibon (2002), p. 133.

# Token-based semantic vector space model for classic poetic Japanese

Xudong Chen<sup>1</sup>, Hilofumi Yamamoto<sup>2</sup>, and Bor Hodošček<sup>3</sup>

## 1 Introduction

This paper explores the effectiveness of the token-based semantic vector space model (Heylen et al., 2012) for describing the classic poetic Japanese vocabulary.

The token-based semantic vector space model represents the semantics of each individual occurrence of a word, while a type-based model aggregates over all occurrences of a word, giving a representation of a word's general semantics (Heylen et al., 2012, p. 17). In type-based models, context- or style-sensitive variation of semantics within word types is averaged and generalized in one vector and thus cannot be described in detail. In token-based models, the description of such variation is possible. Considering the variant referents, meanings, and stylistic usage of the Japanese poetic vocabulary, models on the token level are necessary.

Token-based solutions for the problem of contextualized meanings have been proposed from context-predicting deep learning approaches (e.g. Devlin et al., 2019). Historical data, however, is often too sparse to use the state-of-the-art machine learning methods (Kalouli et al., 2019, p. 109). Another method from a context-counting approach is proposed in Heylen et al. (2012), which does not use any machine learning techniques. Compared to context-predicting deep learning methods, this method is said to have greater transparency (De Pascale, 2019, p. 29). In the present paper, we, therefore, examine the effectiveness of the token-based model for Japanese poetic vocabulary.

## 2 Methods

### 2.1 Materials and preprocessing

Yamamoto and Hodošček (2021a) is used as an annotated corpus of Japanese poetry. Metacodes in Yamamoto and Hodošček (2021b) are used for annotating concept groups of word entries.

We select only poems that are within 41 kana in length. Choka, the long poems, for instance, are excluded. We also exclude any word annotated as a particle, auxiliary, auxiliary verb, prefix, suffix, adverb, interjection, and symbol.

### 2.2 Token-based vectorization

---

<sup>1</sup> Tokyo Institute of Technology

<sup>2</sup> Tokyo Institute of Technology

<sup>3</sup> Osaka University

Suppose that in a corpus with vocabulary size  $d$ , we obtain a token vector of target token  $t$  which occurs in the  $n$ th poem whose number of words is  $l$ . The token vector  $v_{t,n}$  can be calculated with Equation (1),

$$v_{t,n} = \frac{1}{\sum_{i=1}^l w(t, c_i)} \sum_{i=1}^l w(t, c_i) c_i \quad (1)$$

$$c_i = (w(c_i, \text{word}_1) \quad w(c_i, \text{word}_2) \quad \dots \quad w(c_i, \text{word}_d))^T \quad (2)$$

where  $c_i$  is the  $i^{\text{th}}$  context word of word  $t$  in the  $n^{\text{th}}$  poem.  $w(t, c_i)$  is the mutual information between word  $t$  and  $c_i$ .  $c_i$  (Equation (2)) is a vector of context words  $c_i$ , which consists of the mutual information between  $c_i$  and all words in the corpus. For the mutual information, we use PPMI (Bullinaria & Levy, 2007) (Equation (3)),

$$w(a, b) = \text{PPMI}(a, b) = \max\left(0, \log_2 \frac{p(a, b)}{p(a)p(b)}\right) \quad (3)$$

where  $p(a)$ ,  $p(b)$ , and  $p(a, b)$  indicate occurrence probabilities of word  $a$ ,  $b$ , and probability of  $a$  and  $b$  occurring simultaneously, respectively.

### 2.3 Classification task with token vectors

In order to confirm the applicability of the token-based vector space model on Japanese poetic vocabulary, we perform classification tasks with token vectors generated by the method. In the classification tasks, we classify token vectors of two-word pairs and confirm whether 2 words in a pair can be correctly classified.

With metacodes in Yamamoto and Hodošček (2021b), we set the following four different types of classifications:

1. Word pairs matching at the concept group level,  
e.g., flower-flower pair *sakura* (cherry)-*mume* (plum);
2. Word pairs unmatched at the concept group level,  
e.g., flower-bird pair *mume* (plum)-*hototogisu* (cuckoo);
3. Noun pairs with the same lemmas (kana strings), but written with different surface forms,  
e.g., *sakura* (cherry) written as さくら/桜;
4. Verb pairs with the same lemmas, but written in different surface forms,  
e.g., *simu* written as 標む (mark as possession)/染む (dye).

We only include items whose document frequencies are within 20 to 90. Since there are large numbers of type 1 and 2 word pairs, we randomly sample 30 pairs of type 1

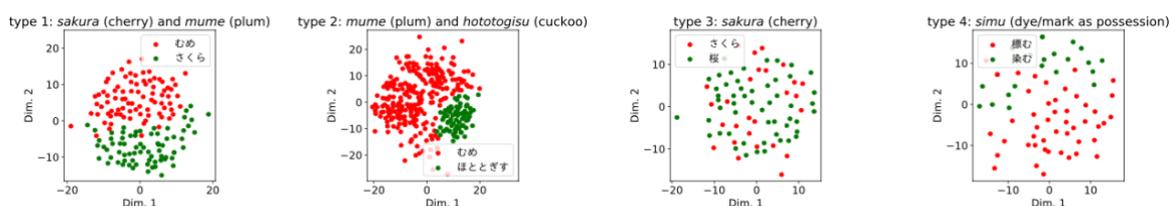
and 2 respectively. In type 3 and 4 pairs, if both surface forms of the type appear in the same poem, we exclude such cases. Finally, we obtained 29 pairs of type 3 and 7 pairs of type 4.

We use logistic regression as a classifier. We randomly sample 20 vectors of each word in each pair and use half of them as training data and the other half as test data.

We use a generalized linear mixed effects model with a binomial distribution to test how the above- mentioned types of word pairs predict the test accuracy (correct number out of 20 test pairs). In the analyses, we include one random effect, the individual differences of the pairs.

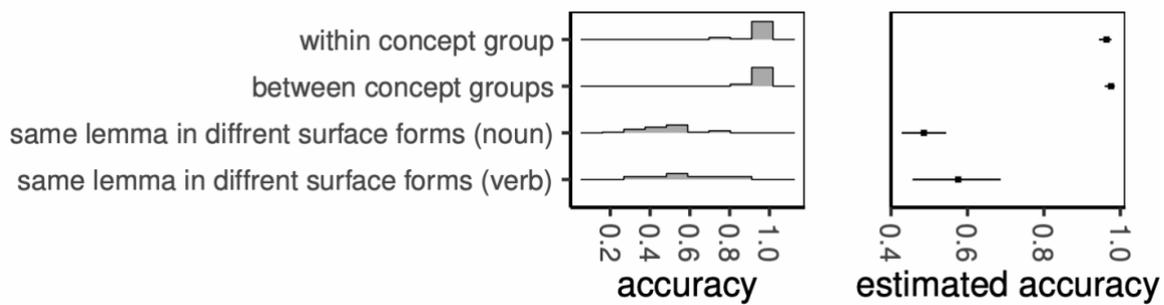
### 3 Results

Four examples from the results are shown in Figure 1. Dimensionality reduction of the vectors is conducted using multidimensional scaling (MDS). The values of the two dimensions in the current paper span a larger range than those reported in Heylen et al. (2012).



**Figure 1: Two-dimensional visualization of token vectors: type 1, 2, and 4 pairs show clear boundaries; the example of type 3 pairs does not show a clear boundary.**

The results of the classification task with token vectors is shown in Figure 2. Test accuracy differs among each type of pairs. Pairs whose lemmas are different (type 1 and 2) have the highest test accuracy, and type 2 does not differ from type 1. Noun pairs with the same lemmas written in different surface forms (type 3) have the lowest accuracy. Verb pairs with the same lemmas written in different surface forms (type 4) also have lower accuracy than those of type 1 and 2. Estimated accuracy of pairs with the same lemmas varies in a larger range than that of pairs in different lemmas.



**Figure 2: Distribution of test accuracy and estimated test accuracy in the classification task: the left shows the distribution of the test accuracy in each type of task; the right shows the accuracy in 95% CI predicted by generalized linear mixed effects model.**

#### 4 Discussion

The vector space generated by the model is sparse. But the token clouds in the 2-dimensional space can reflect the relations among the vocabulary. As shown in Figure 1, token clouds of a pair show clear boundaries when the pair differs more in meanings.

Compared to type 3 and 4, classification in type 1 and 2 pairs has high accuracy. This is because, in most of the cases, word meaning differs more in type 1 and 2 than in type 3 and 4. The accuracy of type 1 classification is slightly higher than that of type 2. This is because pairs in the same concept group share more similar word senses than pairs belonging to different concept groups.

Pairs with the same lemmas cannot be correctly classified. This indicates that information from word types can be important to the current method in a small scale corpus. Most type 3 pairs are often pairs having different surface forms that have no difference in meaning. Therefore the accuracy of type 3 classification is low. On the other hand, there also exist pairs in different surface forms with different meanings in the type 4 pairs (Table 1). The variance of accuracy in type 4 classification is, therefore, greater than that of other types.

**Table 1: Examples of token vectors of type 4 pairs: pairs' surface forms with meaning change are well-classified; pairs' surface forms without meaning change are only correctly classified with a low test accuracy; these cases indicate the importance of contextual information to the pairs with the same word types in classification tasks.**

Type	Surface forms		Test accuracy
	1	2	
<i>okuru</i>	送る (send; see off)	後る (be late)	0.722
<i>koru</i>	懲る (learn a lesson from failures)	樵る (chop down trees)	0.833
<i>simu</i>	染む (dye)	標む (mark as possession)	0.889
<i>kikoyu</i>	聞こゆ (be able to hear)	聞ゆ (be able to hear)	0.500

## 5 Conclusion

We conducted the experiments applying the token-based semantic vector space model (Heylen et al., 2015; Heylen et al., 2012) to Japanese classic poem texts in order to examine the possibilities of the model for small-scale corpora such as the Hachidaishu dataset. We found that although a small corpus generates a sparse vector space, it is possible to observe the differences between words at the token level with token clouds visualization generated by the model. The current method also allows us to relatively successfully classify senses between word pairs.

## References

- Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3), 510–526. <https://doi.org/10/d8pmsm>
- De Pascale, S. (2019). *Token-based vector space models as semantic control in lexical sociolectometry* (PhD Dissertation). KU Leuven. Leuven.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. <https://doi.org/10/ggbwf6>
- Heylen, K., Wielfaert, T., Speelman, D., & Geeraerts, D. (2015). Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua*, 157, 153–172. <https://doi.org/10/gh58qv>
- Heylen, K., Speelman, D., & Geeraerts, D. (2012). Looking at word meaning. An interactive visualization of Semantic Vector Spaces for Dutch synsets. *Proceedings of the EACL-2012 Joint Workshop of LINGVIS & UNCLH: Visualization of Language Patterns and Uncovering Language History from Multilingual Resources*, 16–24.
- Kalouli, A.-L., Kehlbeck, R., Sevastjanova, R., Kaiser, K., Kaiser, G. A., & Butt, M. (2019). ParHistVis: Visualization of Parallel Multilingual Historical Data. *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 109–114. <https://doi.org/10/gh547n>
- Yamamoto, H., & Hodošček, B. (2021a). Hachidaishu part of speech dataset. <https://doi.org/10.5281/zenodo.4835806>
- Yamamoto, H., & Hodošček, B. (2021b). Hachidaishu vocabulary dataset. <https://doi.org/10.5281/zenodo.4744170>

# Open source datasets of the Hachidaishū for the research of classical Japanese poetic vocabulary<sup>1</sup>

Hilofumi Yamamoto<sup>2</sup>, Bor Hodošček<sup>3</sup>

## 1 Introduction

The present paper addresses the curation and publication of an open dataset on Zenodo (<https://zenodo.org/>) for classifying the vocabulary of the Hachidaishū (ca.905–1205). While the dataset was mainly developed in 2009 (Yamamoto 2009), it could not be published until now due to uncertainties in its copyright status. Even if the copyright of the classical text itself has expired, it is still under a reprint copyright which prevents publishing without a clear precedent. In 2016, the National Institute of Informatics (NII) Center of Open Data for the Humanities released the Nijūichidaishū, the Japanese classics dataset created by the National Institute of Japanese Literature (NIJL) under the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license (Kitamoto 2017), which allows us to release our dataset.

Zenodo is a data repository for researchers to store their datasets, founded in 2013 (<https://www.openaire.eu/zenodo-is-launched>) by OpenAIRE (Open Access Infrastructure for Research in Europe) and CERN (European Organization for Nuclear Research). Researchers can upload up to 50GB per dataset, regardless of their research field, and can also cite code, datasets, or things relating to their research that are available on Github.

We will report on the publishing of the Hachidaishū vocabulary dataset, explain the ways in which the dataset will allow researchers to conduct semantic classification of words in the Hachidaishū, and succinctly document the dataset containing part of speech tags and kanji readings.

## 2 Material

The Hachidaishū is a collection of eight anthologies of classical Japanese poetry compiled by the order of Emperors during the 300 years from the Kokinshū (ca.905), the first anthology written in Japanese kana characters, to the Shinkokinshū (1205).(Table 1) The main text is based on a collection of the Nijūichidaishū created by NIJL, and the text is now distributed by both NIJI and NII.(National Institute of Japanese Literature 2016) NIJI provides a poem string search service. NII provides a batch download service for all text

---

<sup>1</sup> This work was supported by KAKENHI, Grant-in-Aid for Scientific Research (C: 18K00528).

<sup>2</sup> Tokyo Institute of Technology

<sup>3</sup> Osaka University

data along with the original images using IIIF (International Image Interoperability Framework). The license for these texts is Creative Commons by SA 4.0 International. The conditions for redistribution are the same.

In addition to the main texts mentioned above, we use, as reference materials, the equivalent data to the Hachidaishū texts contained in the CDROM of the Shimpēn Kokka Taikan, (Shin-pen Kokkataikan Henshū Committee 1996) and those in the Nijūichidaishū database. (Nakamura et al. 1999). We use them as references to attach the readings of Kanji characters to each word.

As mentioned above, since the Hachidaishū consists of Japanese poems published across 300 years, and each collection is published approximately 20 to 80 years apart (42 year intervals on average), it is suitable for research into longitudinal changes in poetic vocabulary.

**Table 1: The list of anthologies in the Hachidaishū: the number of poets in each anthology is based on Shimpēn Kokka Taikan (Kokka Taikan Editorial Committee 1996).**

	name	order	established	editors	poems
1	Kokin	Daigo tennō	ca. 905	Kino Tomonori, Kino Tsurayuki, Ōshikochino Mitsune, Mibuno Tadamine	1,100
2	Gosen	Murakami tennō	ca. 951	Kiyoharano Motosuke, Kino Tokifumi, Ōnakatomino Yoshinobu, Sakanoueno Mochiki Minamotono Shitagō	1,425
3	Shūi	Kazan'in	ca. 1007	Kazan'in	1,351
4	Goshūi	Shirakawa tennō	1086	Fujiwarano Michitoshi	1,218
5	Kin'yō	Shirakawain	ca. 1125	Minamotono Toshiyori	665
6	Shika	Sutokuin	ca. 1151	Fujiwarano Akisuke	415
7	Senzai	Goshirakawain	1188	Fujiwarano Toshinari	1,288
8	Shinkokin	Gotobain	1205	Minamotono Michitomo, Fujiwarano Ariie, Fujiwarano Ietaka, Fujiwarano Masatsune, Jakuren Fujiwarano Sadaie	1,978

### 3 Methods

First, we developed our own dictionary and system to divide the lines of poems of the Kokinshū into units (Yamamoto 2007). We did not accept conjunctions or compound verbs as valid part of speech categories, and adopted the shortest possible unit. In this way, we completed a dictionary that describes the words and concatenation rules of the Kokinshū. After that, the same dictionary was used to divide into units the subsequent collection, the Gosenshū. We checked the words and concatenation patterns that were included in the collection but were not included in the dictionary one by one, and added the missing words and patterns to the dictionary. The same process was repeated, and the dictionary was expanded from the Kokinshū to the Shinkokinshū according to the order of their establishment. In order to prevent the contents added to the dictionary from affecting the unit-divisions processed so far, and to maintain the consistency of the divided units, we

checked each poem processed. If any differences were found with the previously processed results, the differences were output, the content added to the dictionary was reviewed, and the consistency of the dictionary was adjusted. The Hachidaishū Part-of-Speech Dataset was created through the above process.

Because of the variety of notations in Japanese, a metacode was added to indicate the word meaning according to the form in which the word appears. The metacode indicating the lexical system is assigned by referring to the old version of the Word List by Semantic Principles (WLSP) by the National Institute for Japanese Language and Linguistics (NINJAL) (Nakano et al. 1994), and does not correspond to the new version (Asahara 2016, Kato et al. 2018). The classification metacodes are newly added since there are no metacodes for non-independent words such as particles, auxiliary verbs, conjunctions, and proper nouns such as place names (utamakura) and personal names as well as missing words in the WLSP.

#### 4 Construction of two datasets

We will explain the construction of two datasets, the Hachidaishū vocabulary dataset (vocabulary dataset) and the Hachidaishū part-of-speech dataset (POS dataset). Table 2 indicates the construction of the vocabulary dataset.

We will explain the data offset with the first line in Table 2: [01:000001:0007 A00 BG-02-1527-01-0102 47 き 来 < 来 き]. A line consists of 7 columns separated by spaces. The first column “01:000001:0007” consists of 3 fields: 1) anthology ID as indicated in Table2, 2) number of poem, and 3) sequential ID of the token. ID 01 indicates the Kokinshū in this case. The second column indicates the type of token: type A is a single token; type B is a compound token; type C is a breakdown of type B. A00 indicates a single token; A01 indicates a single token and it has another meaning/metacode; B00 indicates a compound token; B01 indicates a compound token which has another meaning/metacode; C00 indicates the first element of the B00/B01.. breakdown; C01 indicates the second element of the B00/B01.. breakdown. The third column “BG-02-1527-01-0102”: classification ID based on semantic categories according to WLSP.(Nakano et al. 1994) The fourth column indicates a POS number used in the morphological analysis system, Chasen.(Matsumoto et al. 2002) The fifth column indicates surface form: a form appears in literary works. The sixth column indicates lemma in kanji writing. The seventh column indicates lemma in kana writing. The eighth column indicates conjugated form in kanji. The ninth column indicates conjugated form in kana.

**Table 2: Data structure of Hachidaishu vocabulary dataset; an example from the Kokinshū Poem #1; left aligned tiny typefaces are for reference and not included in the dataset.**

```

01:000001:0001 A00 BG-01-1630-01-0100 02 年 年 とし 年 とし 年=toshi (year) とし=toshi
01:000001:0001 A10 BG-01-1911-03-1800 02 年 年 とし 年 とし
01:000001:0002 A00 BG-08-0061-07-0100 61 の の の の の=no (particle)
01:000001:0003 A00 BG-01-1770-01-0300 02 内 内 うち 内 うち 内=uchi (inside), うち=uchi
01:000001:0004 A00 BG-08-0061-05-0100 61 に に に に に=ni (particle)
01:000001:0005 A00 BG-01-1624-02-0100 02 春 春 はる 春 はる 春=haru (spring), はる=haru
01:000001:0006 A00 BG-08-0065-07-0100 65 は は は は は は=ha (particle)
01:000001:0007 A00 BG-02-1527-01-0102 47 き 来 く 来 き き=ki (verb: come), 来 (kanji writing of き)
01:000001:0008 A00 BG-03-1200-02-0900 74 に め め に に に=ni (auxiliary verb: perfect), め=lemma of に
01:000001:0008 A10 BG-09-0010-01-0101 74 に め め に に
01:000001:0008 A20 BG-09-0010-03-0200 74 に め め に に
01:000001:0009 A00 BG-09-0010-04-0300 74 けり けり けり けり けり けり=keri (auxiliary verb: past)
01:000001:0010 B00 BG-01-1950-14-0100 02 一 と せ 一 年 ひ と と せ 一 年 ひ と と せ 一年=hitotose (a year), 一 と せ=hitotose
01:000001:0010 C00 BG-01-1950-01-0300 19 一 一 い ち 一 い ち 一=ichi (one), い ち=ichi
01:000001:0010 C01 BG-01-1630-01-0100 02 年 年 とし 年 とし 年=toshi (year), とし=toshi
01:000001:0011 A00 BG-08-0061-10-0100 61 を を を を を を=wo (particle)
01:000001:0012 A00 BG-01-1642-02-0100 02 ご ぞ 去 年 こ ぞ 去 年 こ ぞ 去年=kozo (last year), こ ぞ=kozo
01:000001:0013 A00 BG-08-0061-04-0100 61 と と と と と と=to (particle)
01:000001:0014 A00 BG-08-0065-14-0100 65 や や や や や や=ya (particle)
01:000001:0015 A00 BG-02-3120-01-0100 47 い は 言 ふ い ふ 言 は い は 言ふ=ifu (verb: say), い は=iha (predicative form)
01:000001:0016 A00 BG-03-3012-03-2600 74 ん む む む む ん む む む む ん=n (colloquial form of mu), む=mu (auxiliary verb: inference)
01:000001:0016 A10 BG-09-0010-02-0102 74 ん む む む む ん む む む む
01:000001:0017 B00 BG-01-1641-02-0100 02 こ と し 今 年 こ と し 今 年 こ と し 今年=kotoshi (this year), こ と し=kotoshi
01:000001:0017 C00 BG-03-1000-01-0100 57 こ の こ の こ の こ の こ の
01:000001:0017 C01 BG-01-1630-01-0100 02 年 年 とし 年 とし
01:000001:0018 A00 BG-08-0061-04-0100 61 と と と と と と
01:000001:0019 A00 BG-08-0065-14-0100 65 や や や や や や
01:000001:0020 A00 BG-02-3120-01-0100 47 い は 言 ふ い ふ 言 は い は
01:000001:0021 A00 BG-03-3012-03-2600 74 ん む む む む ん む む む む
01:000001:0021 A10 BG-09-0010-02-0102 74 ん む む む む

```

Table 3 indicates the construction of the POS dataset. We take the Kokinshū, Poem #1 as an example. It is a line, a poem. Tokens are separated by spaces. Each token consists of part-of-speech elements separated by slashes. The first column "10001" contains two elements: the first digit indicates an anthology ID and the rest is a poem ID. The second column and the followings are the information of each token. In the case of nouns and particles, i.e., words that are not conjugated, they are shown in the following format: `text/POS/reading`. In the case of verbs and adjectives, i.e., words that are conjugated, they are shown in the following format: `text/POS:lemma-kanji:lemma-reading/reading`.

**Table 3: Data structure of the Hachidaishu part-of-speech dataset; the first 5 digits are the anthology and poem ID; upper original; lower English translation; \*its POS cannot be determined.**

```

10001 年/名/とし の/格助/の 内/名/うち に/格助/に 春/名/はる は/係助/は き/力変-用:来:</き に/完-用:ぬ:ぬ/に けり/過-終:けり:けり/けり 一とせ/名/ひととせ を/*助/を ごそ/名/こぞ と/格助/と や/係助/や いは/ハ四-未:言ふ:いふ/いは ん/推-終体:む:む/む ことし/名/ことし と/格助/と や/係助/や いは/ハ四-未:言ふ:いふ/いは ん/推-終体:む:む/む

```

```

10001 toshi(year)/noun/toshi no(of)/connecting_particle/no uchi(inside)/noun/uchi ni(indicates_time)/\
case_particle/ni haru(spring)/noun/haru wa(topic)/binding_case/wa ki(come)/kahen_conjugation-conjunctive:\
ku(lemma_kanji):ku(lemma_reading)/ki ni/perfect-conjunctive:nu:nu/ni keru(auxiliary_verb)/\
past-final:keri:keri/keri hitotose(a year)/noun/hitotose wo/case_particle/wo kozo(last year)/noun/kozo \
to/case_particle/to ya/binding_particle/ya iha(say)/yodan_verb-predicative:ifu:ifu/ihana(auxiliary_verb)/\
inference-final:mu:mu/mu kotoshi(this year)/noun/kotoshi to/case_particle/to ya/binding_particle/ya iha(say)/\
yodan_verb-predicative:ifu:ifu/ihana(auxiliary_verb)/inference-final:mu:mu/mu

```

## 5 How to use datasets and access the repository

The POS dataset allows researchers to count the number of tokens for each part of speech; count the number of poems in which the word or part of speech appears; obtain the sequence of words that appear; collect the patterns of the sequence of words; in the case of the vocabulary dataset, to count the number of words in each semantic category, extracting co-occurrence patterns; and so forth. A Jupyter notebook showing examples of Python code needed to conduct some of this is provided in the Github repository alongside the dataset.

Both the vocabulary and POS datasets can be obtained from the Zenodo repository (Yamamoto and Hodošček 2021a,b), and from Github (URL: [url https://github.com/yamagen](https://github.com/yamagen)) as well.

There are advantages to using official repositories like Zenodo and Github: i.e., a DOI specific to the dataset is provided immediately; various bibliographic formats such as Mendeley, BiBTeX, etc. are available; a DOI clarifies the source of the data, ensuring that anyone can use the same dataset and verify the results.

The whole of the dataset is downloadable without user authentication and can be used in the user's preferred environment. Zenodo operates under a license where the data is intended to be downloaded and used. Since it is a portal site type repository, it may be disseminated faster than by publishing it on a personal site. As it is linked with Github, data can be updated and modified and the resulting changes inspected. Also, a DOI is given for each updated version.

## 6 Conclusion

We published two datasets for studying the Hachidaishū vocabulary on Zenodo and Github. We explained how they were created, their structure, how to use them, and introduced the URLs. The copyright issue has been cleared up, and the full text is now available and can be downloaded to promote research on classical Japanese poetic vocabulary at various user levels. For publications using these data, see Chen et al. (submitted to JADH2021). The datasets presented in the current paper are also licensed under the Creative Commons by SA 4.0 International.

## References

- Asahara, Masayuki (2016) "Word List by Semantic Principles (WLSP): a collection of words classified and arranged by their meanings.", <https://github.com/masayu-a/WLSP>.
- Chen, Xudong, Hilofumi Yamamoto, and Bor Hodošček (submitted to JADH2021) "Token-based semantic vector space model for classic poetic Japanese", in *JADH2021 Proceedings of the 11th Conference of Japanese Association for Digital Humanities*.
- Kato, Sachi, Masayuki Asahara, and Makoto Yamazaki (2018) "Annotation of 'Word List by Semantic Principles' Labels for the Balanced Corpus of Contemporary Written

- Japanese”, in *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong: Association for Computational Linguistics.
- Kitamoto, Asanobu (2017) “Center for Open Data in the Humanities (CODH): Activities and Future Plans”.
  - Kokka Taikan Editorial Committee ed. (1996) *Shimpen Kokka-taikan: CDROM Version*: Kadokawa Shoten.
  - Matsumoto, Yuji, Akira Kitauchi, Tatsuo Yamashita, Osamu Imaichi, and Tomoaki Imamura (2002) *Morphological Analysis System ChaSen Version 2.2.9 Manual*, Nara Institute of Science and Technology.
  - Nakamura, Yasuo, Yoshihiko Tachikawa, and Mayuko Sugita (1999) *Kokubungaku kenkyū shiryōkan dētabēsu koten korekushon “Niju ichidaishu” Shōhobanbon CD-ROM (Database Collection by National Institute of Japanese Literature “Niju ichidaishu” the Shōho edition CD-ROM)*: Iwanami Shoten.
  - Nakano, Hiroshi, Ooki Hayashi, Hisao Isii, Makoto Yamazaki, Masahiko Ishii, Yasuhiko Kato, Tatuō Miyazima, and Akio Tsuruoka (1994) *Bunrui goi hyō furoppī ban (Word List by Semantic Principles, floppy disk version)*, Vol. 5 of Kokuritsu Kokugo Kenkyūjō gengo shori data shū (National Language Research Institute language data), Tokyo: Dainippon Tosho.
  - National Institute of Japanese Literature (2016) “The Niju ichidaishu Japanese Classics Dataset”, <http://codh.rois.ac.jp/pmjt/book/200007092/>, <http://kotenseki.nijl.ac.jp/biblio/200007092/>.
  - Shin-pen Kokketaikan Henshū Committee ed. (1996) *Shimpen Kokka-taikan: CDROM Version* : Kadokawa Shoten.
  - Yamamoto, Hilofumi and Bor Hodošček (2021a) “Hachidaishu part of speech dataset”, <https://doi.org/10.5281/zenodo.4835806>.
  - Yamamoto, Hilofumi and Bor Hodošček (2021b) “Hachidaishu vocabulary dataset”, <https://doi.org/10.5281/zenodo.4744170>.
  - Yamamoto, Hilofumi (2007) “Waka no tame no Hinshi tagu zuke shisutemu / POS tagger for Classical Japanese Poems”, *Nihongo no Kenkyū / Studies in the Japanese Language*, Vol. 3, No. 3, pp. 33–39.
  - Yamamoto, Hilofumi (2009) “Thesaurus for the Hachidaishu (ca.905–1205) with the classification codes based on semantic principles”, *Nihongo no Kenkyū / Studies in the Japanese Language*, Vol. 5, No. 1, pp. 46–52.

# Exploring Metadata Quality Issues in Non-English Corpora: Preliminary Assessments of HathiTrust Records of Late Imperial Chinese Books

Wenyi Shang<sup>1</sup>, Jacob Jett<sup>2</sup>, J. Stephen Downie<sup>3</sup>

## Introduction

Large online bibliographic datasets have brought about new possibilities for studies of late imperial Chinese books, but the tension between Anglo-American cataloging practices and Chinese books could cause vital problems. In this poster we conduct a preliminary examination of this problem through a case study of HathiTrust<sup>4</sup> (HT) MARC metadata records describing the Chinese books published in the 16th, 17th, and 18th century by quantitatively analyzing them. Referencing the time scope used by Chartier (1983) in the history of Western written culture, a set of Chinese books published in the 16th, 17th, and 18th centuries were collected from the HT.

## Methods

A workset of all books published between 1500 and 1799 and which were written in Chinese language in HT's collection was built using the "Workset Builder 2.0 for Extracted Features 2.0,"<sup>5</sup> which is a tool designed by the HathiTrust Research Center (HTRC)<sup>7</sup> that allows users to extract features from unigram queries over the HT corpus. We used the following query for constructing the workset for our analysis: "(pubDate\_t:15\*\* OR pubDate\_t:16\*\* OR pubDate\_t:17\*\*) AND language\_t:chi." The HT identifiers of all works satisfying the query were collected (9,437 books in total). Next, the MARC records<sup>8</sup> describing each book was collected using the HathiTrust Bibliographic API via their HT identifiers. The results were then analyzed and visualized with the programming language Python. The MARC records (in XML format) were first parsed with the Python library "Beautiful Soup." Next, the parsed data were analyzed with the Python libraries "Pandas" and "Regex" (regular expression) and were visualized with

---

<sup>1</sup> School of Information Sciences, University of Illinois at Urbana-Champaign

<sup>2</sup> School of Information Sciences, University of Illinois at Urbana-Champaign

<sup>3</sup> School of Information Sciences, University of Illinois at Urbana-Champaign

<sup>4</sup> <https://www.hathitrust.org>

<sup>5</sup> <https://solr2.htrc.illinois.edu/solr-ef>

<sup>6</sup>

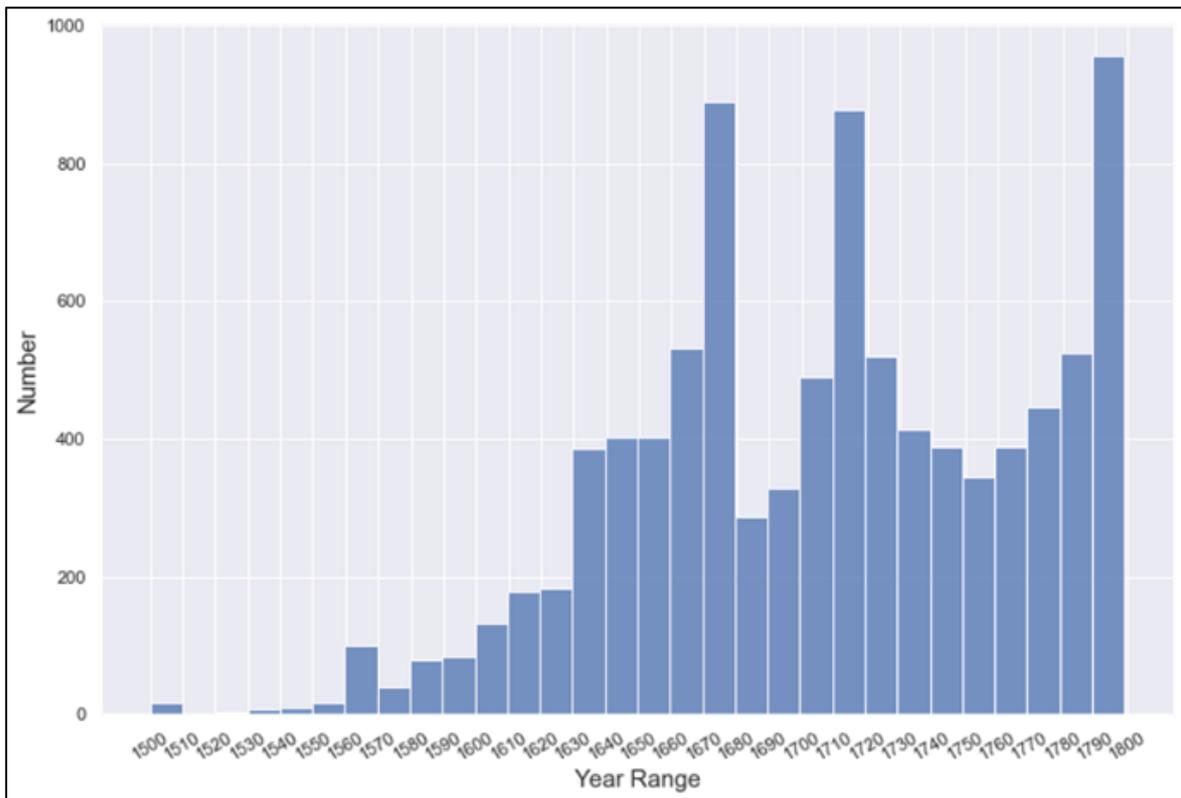
<https://wiki.htrc.illinois.edu/display/COM/HTRC+Workset+Builder+2.0+%28Beta%29+for+Extracted+Features+2.0>

<sup>7</sup> <https://analytics.hathitrust.org>

<sup>8</sup> <https://www.loc.gov/marc/marcinf.html>

<sup>9</sup> <https://www.loc.gov/marc/bibliographic>

the Python libraries “Matplotlib” and “Seaborn.” These analyses allowed us to develop a baseline of informative data describing features of the selected works.



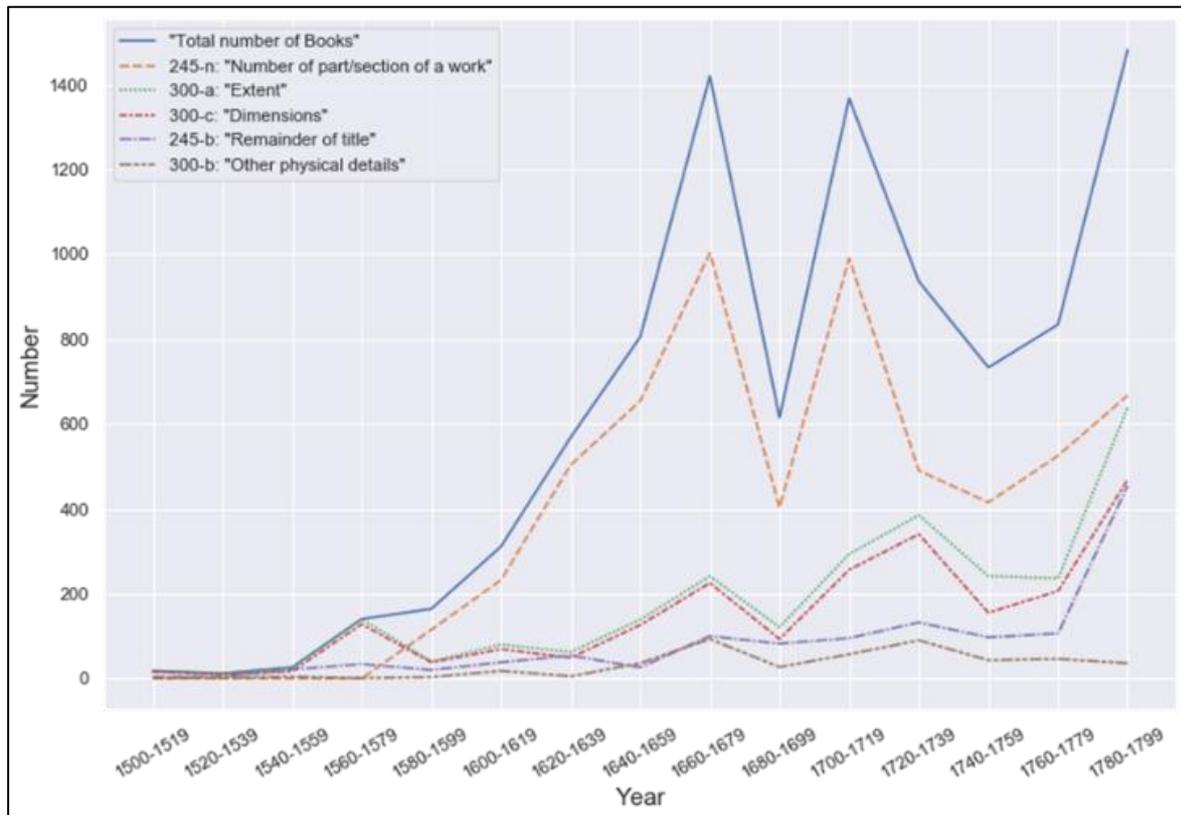
**Figure 1: Chinese Books Published in Different Year Ranges (in decades) in HathiTrust Digital Library**

## Findings

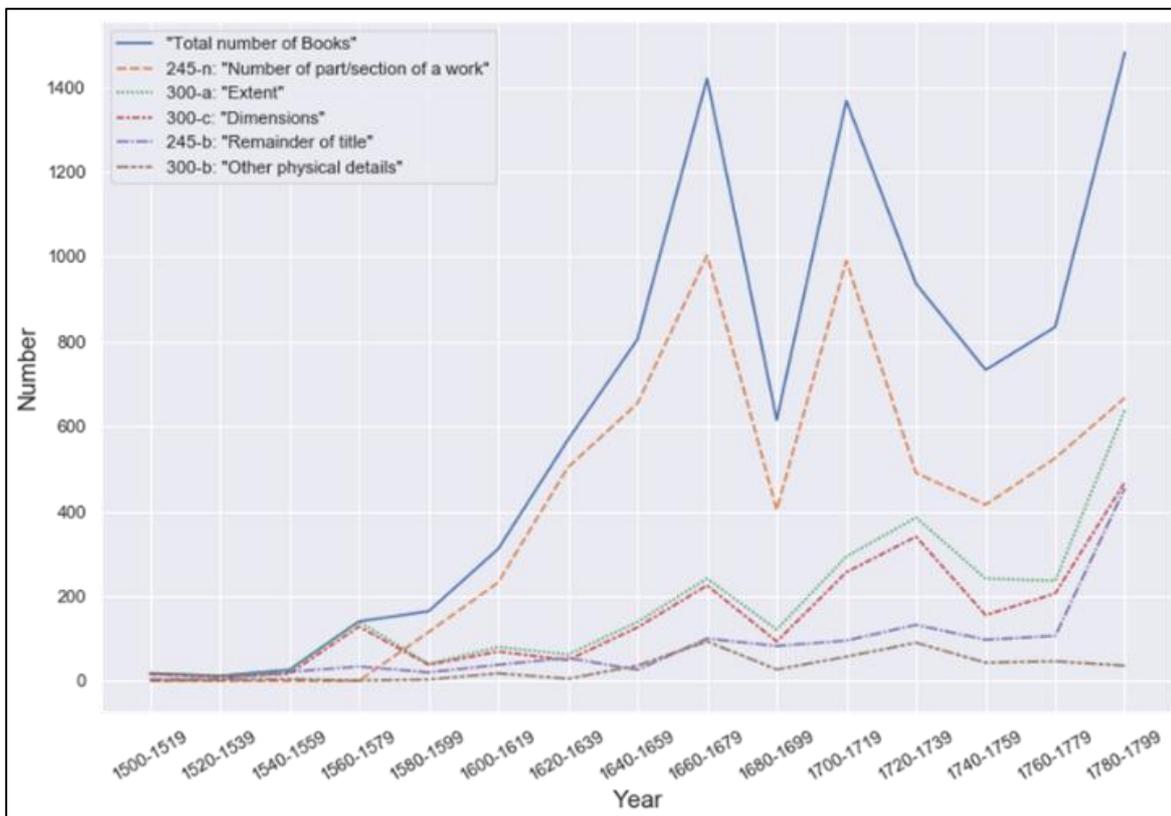
Several of the figures showcase results that were unexpected. For instance, Figure 1 does not display an expected reduction in printing activities during the height of the Manchu conquest in the 1640s and 1650s. Instead the metadata from the HT corpus demonstrates an upward trending progression. Similarly, while it would be natural to attribute the valley between the 1720s and the 1780s to the severe situation of literary inquisition<sup>10</sup> during the reign of the Qianlong Emperor (r. 1735–1796), when 53 cases of literary inquisition were recorded (Wong, 2000), a similar explanation is not suitable for the unexpected valley occurred between the 1680s and the 1700s, during the reign of the Kangxi Emperor (r. 1661–1722), when literary inquisition was relatively uncommon.

<sup>10</sup> The literary inquisition is the official persecution of intellectuals for their writings in imperial China, which reached its peak in the Qing dynasty (1644–1912). The imperial government freely censored everything that was written and convicted the intellectuals for any writings that the ruler considered offensive.

Furthermore, as is shown in Figure 2, the change of the number of books containing the MARC field 245-n “number of part/section of a work” across different time periods does not follow a similar trend as numbers of books containing other fields related to physical descriptions (e.g., from 1700–1719 to 1720–1739). The field 245-n is used as an alternative rather than a supplement to the other fields related to physical descriptions. Another significant problem with the MARC record of late imperial Chinese books is its lack of genre information. As shown in Figure 3, only a very small proportion of books contain such information (out of the total 9,430 books, only 90 contain the field 650-v).



**Figure 2. Number of Chinese Books Containing Fields Related to Physical Descriptions in HathiTrust Digital Library**



**Figure 3. Number of Chinese Books Containing Fields Related to Genre Information in HathiTrust Digital Library**

Overall, several potential problem areas were noticed. These included:

- [1] Same data in multiple fields (245-n and 300-a “extent”; 100-a “personal name” and 245-c “statement of responsibility, etc.”).
- [2] Lack of normalization (245-n).
- [3] Transcribed from source without translation and explanations (260-a “place of publication, distribution, etc.”).
- [4] Lack of essential data (650-a “topical term or geographic name entry element”; 650-v “form subdivision”; 650-x “General subdivision”; 650-y “chronological subdivision”; 650-z “geographic subdivision”).

## Conclusion

These results show that, despite the existence of established guidelines for cataloguing ancient Chinese books (Research Libraries Group, 2000), various significant problems are still prevalent in the HathiTrust MARC metadata records of late imperial Chinese books. Among these problems is that the guideline hasn’t been strictly followed in all cases. This poster calls attention to the burgeoning need for retrospective cataloging. Therefore, although there are many exciting possibilities opened up by digital methods on humanities studies, and “find or organize works” is arguably one of “seven ways humanists are using computers to understand text,” (Underwood, 2015) one must be very cautious with regards to the results yielded by distant reading of large-scale metadata describing non-Western books at this point.

During the next phase of this work, we will be collecting a comparable workset comprising an equal number of English-language works from the same time period to analyze whether or not the metadata quality issues observed in relation to the Chinese-language workset are unique to it.

## Reference

- Chartier, R. (ed.) (1989) *The Practical Impact of Writing*. In Goldhammer, A. (tran.), *A History of Private Life, Volume III: Passions of the Renaissance*. Cambridge, MA and London, UK: The Belknap Press of Harvard University Press, pp. 111–159.
- Research Libraries Group (2000) *Cataloging Guidelines for Creating Chinese Rare Book Records in Machine-Readable Form*. Mountain View, CA: Research Libraries Group.
- Underwood, T. (2015) *Seven Ways Humanists Are Using Computers to Understand Text, The Stone and the Shell*, 4 June. <https://tedunderwood.com/2015/06/04/seven-ways-humanists-are-using-computers-to-understand-text/> (accessed 13 December 2020).
- Wong, K. C. (2000) *Black's Theory on the Behavior of Law Revisited IV: the Behavior of Qing Law*, *International Journal of the Sociology of Law*, 28(4): 327–374.

# Dataset Construction for Cross-genre Plot Structure Extraction

Hajime Murai<sup>1</sup>, Shuuhei Toyosawa<sup>1</sup>, Takayuki Shiratori<sup>1</sup>, Takumi Yoshida<sup>1</sup>, Shougo Nakamura<sup>1</sup>, Yuuri Saito<sup>1</sup>, Kazuki Ishikawa<sup>1</sup>, Sakura Nemoto<sup>1</sup>, Junya Iwasaki<sup>1</sup>, Akiko Uda<sup>1</sup>, Shoki Ohta<sup>1</sup>, Arisa Ohba<sup>1</sup>, Takaki Fukumoto<sup>1</sup>

## Introduction

Several studies have long established that it is possible to extract the common plot structure of specific genre stories when many specific genre stories are collected [1-3]. Based on these old humanistic studies, recent research focusing on several specific genre stories has clarified that the quantitative and objective extraction of common plot structures can be executed using computational methods [4, 5]. In these recent studies, the plot structures were described as sequences of symbolized scenes or functions. The common plot structures of specific genres were extracted using quantitative methods for those symbolized sequences.

However, these past studies focused only on specific genres and thus the common characteristics of general plot structures have not been recognized. The present study is the first to develop a common symbol set for describing the plot structures of several different genres. The identification of symbols that are common between different story genres enables the comparison of the characteristics of each story genre. These symbol sets can be utilized for extracting common patterns of general stories. Moreover, the extracted common patterns could become the foundation for automatic story generation systems.

## Target contents

To compare different story genres, several popular genres in modern Japanese entertainment culture were selected on the basis of comic and game sales rankings. The selected genres were “Adventure,” “Battle,” “Love,” “Detective,” and “Horror.” To extract typical plot structures for each genre, works of combined genres (such as “love comedy”) were eliminated, and popular short stories were picked up based on sales rankings. If there were not enough popular short stories, popular long stories were divided into short stories based on the changes in the purpose of the protagonists of the stories [6]. Subsequently, the selected stories were divided into plot elements (scenes), and the categories were inductively constructed manually. Table 1 shows the final 29 categories of the plot elements. This category table and categorized plot data were verified by several analysts.

---

<sup>1</sup> Future University Hakodate

Details of each genre, that is, number of stories analyzed, number of plot elements found, and average length of stories, are shown in Table 2. Based on the analysis, a total of 873 stories in five genres were divided into 7695 plot elements and categorized into 29 types.

### **Differences and characteristics for each genre**

To extract differences between genres, a chi-square test residual analysis was performed. The results are presented in Table 3.

Table 3 shows the differences between and characteristics of each genre. For instance, the plot elements about “travel route” frequently appeared in “Adventure,” and the elements related to human relationships frequently appeared in “Love.” These characteristics correspond to a general understanding of each genre.

### **Conclusions and future work**

In this study, a common symbol set for the categorization of the plot structure in several different story genres was developed. Moreover, it was established that the characteristics of each genre can be quantitatively extracted.

Using the method used in this study, future studies could extract frequently appearing patterns in each genre. Furthermore, the findings of the study can act as the foundation for automatic story generation algorithms for general stories.

### **Reference**

- [1]. **Barthes, R.**, (1968). *Elements of Semiology*. Hill and Wang, New York, USA.
- [2]. **Propp, V.**, (1968). *Morphology of the Folk Tale*. U of Texas P, USA.
- [3]. **Campbell, J.**, (1949). *The Hero with a Thousand Faces*. Pantheon Books, USA.
- [4]. **Murai, H.**, (2014). “Plot Analysis for Describing Punch Line Functions in Shinichi Hoshi’s Microfiction”, 2014 Workshop on Computational Models of Narrative, (Eds. Mark A. Finlayson, Jan Christoph Meister, and Emile G. Bruneau), *OpenAccess Series in Informatics*, 41:121-129.
- [5]. **Murai, H.**, (2020). “Factors of the Detective Story and the Extraction of Plot Patterns Based on Japanese Detective Comics”, *Journal of the Japanese Association for Digital Humanities*, 5(1): 4-21.
- [6]. **Nakamura, S. and Murai, H.**, (2020). “Proposal of a method for analyzing story structure of role-playing games focusing on quests structure”, *Computer and Humanities Symposium*, 2020: 149-156 (in Japanese).

**Table 1: Categories of plot elements**

Arrival	Encounter with the protagonist, including events such as birth and revival
Leaving	Leaving from the story, including permanent leaving such as death
Change	Change in a character's attributes (e.g., swap, transform, and face change by plastic surgery)
Ability improvement	Positive change in a character's ability
Ability decline	Negative change in a character's ability
Getting travel route	A character is able to move
Escape	Escaping from something (e.g., retreat, withdrawal, liberation, and prison break)
Losing travel route	A character cannot move (e.g., losing transportation facilities, detention, kidnapping, arrest)
Search	Effort for obtaining information (e.g., exploration, survey, and research)
Discovery	Disclosure of some information or hidden truth
Misunderstanding	A character has a misunderstanding
Doubt	A character notices something suspicious and has doubts
Concealment	Some scenes about hiding information (e.g., concealment, disguise, scam)
External information	External information presentation for audiences through elements such as prologue and epilogue to explain about the world of the story
Order, promise	It includes not only promise, transaction, and compliance, but also warning and prophecy.
Violation	It includes crime, negligence, ignorance of warnings, and inattention.
Intention, request	It includes scenes related to characters making decisions, that is, scenes involving wishing, request, persuasion, and invitation.
Completion of request	A scene that mainly consists of fulfilment of a request
Failure of request	A scene that mainly consists of a failure or refusal to grant or fulfil a request
Insanity	Situation wherein the character cannot control himself/herself (e.g., madness, confusion, and possession by evil spirits)
Positive relationship	Positive changes in human relationships (e.g., conversion, reflection, reconciliation, expression of gratitude)
Negative relationship	Negative changes in human relationships (e.g., quarrel, betrayal, arrogance, disgust)
Positive love relationship	Positive changes in human love (e.g., falling in love, confession of feelings, date, marriage)
Negative love relationship	Negative changes in human relationships in the context of love (e.g., jealousy, broken heart, divorce)
Aid	It includes many types of "help," such as rescue, nursing, assistance, encouragement, and sacrifice.
Interference	It includes not only explicit interferences but also acts that intentionally make the other person uncomfortable.
Confrontation	Combat and competitions, including sports
Everyday	Scenes of ordinary everyday life
Disaster	It includes not only natural disasters, but also accidents and mental crises such as severe depression

**Table 2: Data analyzed for each genre**

	Stories	Plot elements	Average length
--	---------	---------------	----------------

Adventure	206	1795	8.7
Battle	243	2023	8.3
Love	123	1119	9.1
Detective	134	1271	9.5
Horror	167	1487	8.9

**Table 3: The prevalence of the identified plot elements in each genre**

	Adventure	Battle	Love	Detective	Horror
Arrival	▽▽55	▽132	88	▲▲138	▲▲170
Leaving	▽▽47	▲▲141	▽▽16	▽▽33	▲▲144
Change	▲▲19	6	3	▽1	7
Ability improvement	▲▲119	▲▲134	▽▽6	▽▽0	72
Ability decline	▽33	▲▲89	▽▽14	▽▽11	▲49
Getting travel route	▲▲106	52	▽▽0	▽▽1	35
Escape	39	▽26	▽11	32	▲40
Losing travel route	▲▲91	▽▽29	▽▽4	▲▲86	▽▽25
Search	▲▲134	▽▽30	▽▽4	▲▲172	▽▽40
Discovery	▽▽234	▽▽182	▽▽84	▲▲444	227
Misunderstanding	9	▽▽5	▲▲26	4	9
Doubt	50	68	▽▽13	▽▽26	▲▲106
Concealment	▽▽7	27	23	▲▲36	29
External information	▽▽20	▽▽18	28	▽▽5	▲▲74
Order, promise	22	33	12	▽▽0	▲▲38
Violation	▽6	▲▲24	10	▽▽0	14
Intention, request	▲▲196	163	▽▽56	▽81	113
Completion of request	▲▲68	▽24	▽▽4	▽▽1	33
Failure of request	8	10	3	▽0	4
Insanity	18	16	▲▲17	▽▽0	7
Positive relationship	▲▲59	40	▲▲39	▽▽9	▽▽6
Negative relationship	▽▽12	▽▽5	▲▲348	▽▽1	▽▽7
Positive love relationship	▲▲62	62	▲40	▽▽12	▽▽24
Negative love relationship	▽▽1	▽▽1	▲▲117	▽▽0	▽▽2
Aid	106	▲▲155	▲▲78	▽▽22	▽▽26
Interference	▽▽75	▲▲251	▽▽34	▽▽55	▽▽71
Confrontation	▲▲150	▲▲229	▽▽3	▽▽34	▽▽10
Everyday	▽▽2	▽▽26	20	▲▲51	▲▲65
Disaster	47	44	16	▽16	40

▲▲ Large in 1% statistical significance      ▽▽ Small in 1% statistical significance  
 ▲ Large in 5% statistical significance      ▽ Small in 5% statistical significance

## Basic Plot Structure in the Adventure and Battle Genres

Yuuri Saito<sup>1</sup>, Takumi Yoshida<sup>1</sup>, Shougo Nakamura<sup>1</sup>, Kazuki Ishikawa<sup>1</sup>, Shoki Ohta<sup>1</sup>,  
Arisa Ohba<sup>1</sup>, Takaki Fukumoto<sup>1</sup>, Hajime Murai<sup>1</sup>

### Background

There have been many studies on extracting the basic structure of stories. Propp found that the structure of Russian magical folktales consisted of a combination of 31 different functional structures [1]. Similarly, Shigehisa conducted a study of structural analysis and quantification methods for video works by applying analytical methods for plot patterns related to "myths and folktales." [2] In another study, Nakamura proposed a method for analyzing plot patterns in role-playing games by focusing on quest structures [3]. There have also been studies that analyzed role-playing games [4] and studies that extracted the basic narrative structure of Shinichi Hoshi's short stories [5]. Although all these studies focused on specific single genres, we believe that if we isolate an element of the narrative structure, we can discover similar plot patterns in other narrative genres.

In this study, after analyzing the narrative structure of multiple narrative genres, we created categories that represent narrative functions common to the genres and used these categories to extract the basic narrative structure for each genre.

In this paper, we will explain the results of the analysis, discuss their implications, and present a comparison of plot structure from the adventure and battle genres.

### Object and method of analysis

The works to be analyzed were selected by referring to the total number of copies of Japanese comics and the number of games sold. The genres analyzed were "Adventure," "Battle," "Romance," "Detective," and "Horror." In the "Adventure" category, the analysis focused mainly on role-playing games such as "Pokemon" [6] "Dragon Quest" [7] and "Final Fantasy" [8] In the "Battle" category, "Dragon Ball" [9] (Vol. 1-17, comic's episodes 1-194), "Knights of the Zodiac" [10] (Vol. 1-15, comic's episodes 1-194), and "Bleach" [11] (Vol. 1-21, comic's episodes 1-181) were analyzed. In order to extract a typical narrative structure, the full-length works were divided into quest units according to the goals and objectives that guide the main character's actions [3].

Subsequently, the work was further divided by scenes, which were classified into categories according to their functions in the story. In this paper, we assign tag for one function to one scene. Table 1 shows a list of the 29 categories that were created. Table 2

---

<sup>1</sup> Future University Hakodate

and Table 3 show the number of quests and scenes in each analysis of the "Adventure" and "Battle" categories, respectively, and the average quest length, which indicates how many functions a quest could be divided into on average.

**Table 1: Category for plot elements.**

Arrival	Encounter with the protagonist, including such as birth and rival.
Leaving	Leaving from the story, including permanent leaving such as death.
Ability improvement	Good change for a character's ability.
Ability decline	Bad change for a character's ability.
Getting travel route	A character becomes to be able to move.
Search	Effort for obtaining information, such as exploration, survey and research.
Discovery	Disclosure of some information or hidden truth.
Intention, request	It includes scenes related to characters decision making, such as wishing, request, persuasion, and invitation.
Positive relationship	Positive changes of human relationship such as conversion, reflection, reconciliation, and thanks.
Negative relationship	Negative changes of human relationship such as quarrel, betrayal, arrogance and disgust.
Aid	It includes many types of 'help', such as rescue, nursing, assistance, encouragement, and sacrifice.
Interference	It includes not only explicit interferences but also some acts that intentionally make the other person uncomfortable.
Confrontation	Combats and competitions. Including sports.

**Table 2: Analysis results for "Adventure".**

	Quest	Function Average	Average Quest Length
Pokemon	21	182	8.7
Dragon Quest	27	771	28.6
Final Fantasy	158	842	5.3
RPG total	206	1795	8.7

**Table 3: Analysis results for "Battle".**

	Quest	Function Average	Average Quest Length
Dragon Ball	165	533	3.2
Knights of the Zodiac	61	1245	20.4
BLEACH	17	245	14.4
Battle Total	243	2023	8.3

### Analysis Results and Discussion

For the extraction of narrative structure for each genre, we chose the method of taking 4-grams for each scene and combining the most frequent ones in order from the top to the bottom to include them as typical plot patterns that frequently appear in the genre. Therefore, the resulting plot patterns encompass the continuous patterns with high

frequency of occurrence in the target narrative genre. Figure 1 shows some of the typical plot patterns and their contents that frequently appear in "Adventure" and "Battle." The bracketed areas occupy the places where both the upper and lower patterns are possible.

In "Adventure," the early part tended to feature events that opened up the way to explore different places. In the middle, there were frequent occurrences of events where requests are made to the main character, the character makes a decision, and the hidden truth is disclosed. It ended with a battle against the enemy, victory, growth of the main character, and the acquisition of gold and silver treasures.

In "Battle," after an event in which a character appears for the first time in the early stages, the protagonist often decides to fight after being interrupted by an enemy or is interrupted by an opponent who decides to fight. In the middle, after an event in which the protagonist helps an ally or is helped by an ally, or an event in which a hidden truth is disclosed, there is often a battle with the enemy. At the end, while there is one event in the middle of the battle where the player is in trouble, it often ended with the player's ability improving and them winning the battle against the enemy or growing after the victory over the enemy.

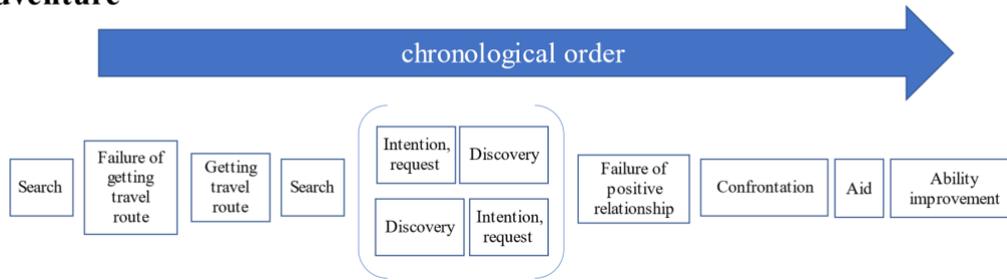
The common features of the two genres are that the main character's objective is often declared in the early stages of the story, the hidden truth is often disclosed in the middle of the story, battles tend to occur from the middle to the end of the story, and ability improvement tends to occur near the end of the story. One difference is that in "Battle," events such as disturbances occur early in the story and objectives that force the protagonist to fight the enemy are more likely to be shown, while in "Adventure," exploratory events are more likely to occur frequently. The other point is that in "Battle," there is an emphasis on staging that places serious obstacles or challenges in the protagonist's path, such as "Ability decline".

## **Conclusion**

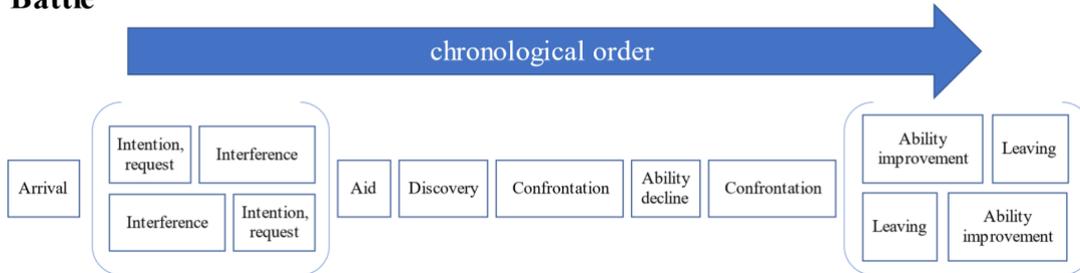
In this study, we created categories that can classify common narrative functions among genres, and using these categories, we extracted typical plot patterns that frequently appear in each genre. As a result of discussing and comparing them, we found that "Adventure" and "Battle" share the common feature wherein "ability improvement" and "battle" tend to occur at the end of the story.

As a future task, we would like to increase the number of works to be analyzed and collect and analyze more data, so that we can continue to improve the common categories and plot patterns to make them more general and typical. We would also like to make these data applicable to automatic story generation.

## Adventure



## Battle



**Figure 1: Typical plot patterns with frequent occurrences of "Adventure" and "Battle".**

## Reference

- [1]. Propp, V., (1968) Morphology of the Folk Tale. U of Texas P, USA.
- [2]. Shigehisa, R., Kida, S. and Takada, A., (2007). "A study of quantification methods for analyzing the structure of video works", In Proceedings of the 69th national conference of the Japan Society of Information and Knowledge, 2007, 499-500 (in Japanese).
- [3]. Nakamura, S. and Murai, H., (2020). "Proposal of a method for analyzing story structure of role-playing games focusing on quests structure", Computer and Humanities symposium, 2020: 149-156 (in Japanese).
- [4]. Ooki, T., Nishijima, K., Uchida, A. and Takada, A., (2006). "An application of narrative structure analysis to RPG analysis", In Proceedings of the 68th national conference of the Japan Society of Information and Knowledge, 2006, 497-498 (in Japanese).
- [5]. Murai, H., Matsumoto, N., Sato, C. and Tokosumi, A., (2011). "Towards the numerical analysis of narrative structure: The characteristics of narrative structure within the short-short stories of Shinichi Hoshi", Journal of the Japan Society of Information and Knowledge, (2011): 6-17 (in Japanese).
- [6]. Pokemon RED • BLUE, Nintendo Co., Ltd., 1996
- [7]. Dragon Quest, SQUARE ENIX HOLDINGS CO., LTD., 1986.
- [8]. Final Fantasy, SQUARE ENIX HOLDINGS CO., LTD., 1987.
- [9]. Dragon Ball, Akira Toriyama, SHUEISHA, 2002.
- [10]. Knights of the Zodiac, Masami Kurumada, SHUEISHA, 1997.
- [11]. Bleach, KUBO TAITE, SHUEISHA, 2016.

# Construction of ShiJi Spatiotemporal Information Platform on the Framework of Research-oriented Knowledge Bases

Jung-Yi Tsai<sup>1</sup>, Pi-Ling Pai<sup>1</sup>, Hsiung-Ming Liao<sup>1</sup>, You-Jun Chen<sup>2</sup>, Richard Tzong-Han Tsai<sup>13\*</sup>, I-Chun Fan<sup>2</sup>

## Introduction

Developed by the Center for GIS at Academia Sinica, the ShiJi Spatio-Temporal Information Platform is an integrated system to present historical information about people, places, relations, and events in Records of the Grand Historian known by its Chinese name ShiJi, in its spatial and temporal context. Given that the narrative of ShiJi provides insightful information outlining the causality and spatio-temporal characteristics of historical events, ShiJi has served as an essential reference for researchers and scholars nowadays to clarify the chronological and geographical evolution of related events in ancient times. Therefore, this research uses the text of ShiJi as the material for event analysis and applies the events in ShiJi into a time-space integrated information system. For the compilation of research-oriented knowledge bases, the platform is built upon linked data infrastructure, committed to the availability and accessibility of historical, geographical, and topological data and research results for the benefit of researchers, scholars, and students.

## Datasets

The platform integrated multiple related historical materials and maps that illustrate historical events recorded in ShiJi, referred to the interpretation of historical experts, and extracted information of time and space in the texts. The platform comprises three primary datasets: texts in ShiJi with supplementary texts from related historical materials, historical maps made by professor Panqing Xu (Xu, 2010), and the database Chinese Civilization in Time and Space (CCTS) (Academia Sinica, 2002).

The texts in ShiJi are reorganized from a series of biographies into 360 main sections and around 1200 events in chronological order, spanning over 2500 years from the ancient Yellow Emperor to the Han Dynasty. The source of an event text may come from multiple chapters in ShiJi, composed in a contextual chronological narrative, and can also

---

<sup>1</sup> Center for GIS, Research Center for Humanities and Social Sciences, Academia Sinica, Taiwan

<sup>2</sup> Institute of History and Philology, Academia Sinica, Taiwan

<sup>3</sup> Department of Computer Science and Information Engineering, National Central University, Taiwan

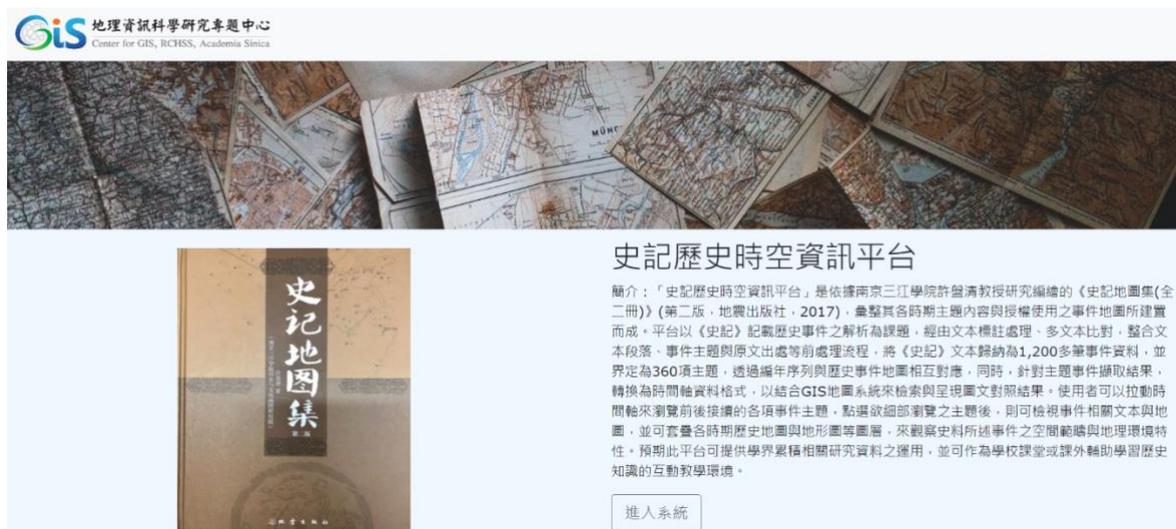
\* corresponding author

be cross-compared with other historical documents such as Hanshu, Zizhi Tongjian, and then integrated into respective event texts. The event texts are accompanied by corresponding maps created by digitizing and georeferencing and then overlaying or aligning them on shaded relief and street maps provided by public web map tile service. A total of 224 event maps have linked to the event texts on the Shiji platform.

For the 6872 place names identified in the event texts on the GIS-based map interface, we further carried out named entity linking corresponding to CCTS for coordinate positioning. By referring to the historical place name database of CCTS, it is helpful for subsequent integration with the infrastructure of CCTS as a basis for dynamically marking the locations in the event texts on the multi-period historical map layers and conducting related research applications.

## System Design and Examples of Usage

To introduce the characteristics of the data used by the platform, we first created an instruction page (Figure 1), presented before users enter the core system.



**Figure 1: Home Page of the ShiJi Spatio-Temporal Information Platform**

The datasets mentioned above are interrelated, so we design a data interface (DATA API) to access the relevant data of the historical events and integrate the maps and text data on the platform interface (Figure 2). First, the platform will access the chronological event database to form a historical event timeline with event titles on the webpage. When users click the box on the historical event timeline to access an event, the related event text will display in the middle of the page. Secondly, as shown by arrow 2 in Figure 2, the map data access is mainly responsible for listing the event maps currently integrated with the event. The historical coordinates are responsible for displaying the spatial information of the place names in the text. When the user clicks on the place names

(highlighted in yellow) in the text, the system will query the historical coordinates database (arrow 3 in Figure 2), return the coordinates, and then display points on the GIS-based map interface as shown in the red box in Figure 2.



**Figure 2: System Architecture**

The platform mainly provides two aspects of the information: (1) the continuous development of events, (2) the combination of events and geographical information. Users can follow a given history timeline and observe event developments. For example, as shown in Figure 3, the event "吳起死 (Wu Qi died)" is followed by subsequent occurrences such as "齊因韓與秦、魏戰而襲燕取國之桑丘 (Because of Han's war with Qin and Wei, Qi took advantage of the situation to seize the land of Yan, Sanqiu)" and "田氏併齊與威王因齊立 (The Tian family annexed and inherited the country of Qi and had Tian Yinqi as the king of Qi)". Users can further analyze the subsequent impact on the powers and figures of various countries after an incident has occurred.

In addition, users can take geographical information into account while analyzing events. As shown in Figure 4, in the event "齊因韓與秦、魏戰而襲燕取國之桑丘", users can observe from the historical map and analyze whether the army's movement was affected by the topography.



transformed over time. The platform also indicates the research potential of linking textual data with map data, facilitating a grander scale of digital humanities research. In the future, we will use NLP technology to identify relationships between people and event categories. By integrating more diverse data into this platform, events can be analyzed from a more well-rounded perspective.

### **Reference**

- [1]. Xu, P. -Q. (2010). *Atlas of ShiJi*. Beijing: Seismological Press.
- [2]. Academia Sinica. (2002). Chinese Civilization in Time and Space (CCTS), First Edition, Taipei. Website: <https://cts.sinica.edu.tw/>

# Cross-genre Plot Analysis of Detective and Horror Genres

Junya Iwasaki<sup>1</sup>, Shuuhei Toyosawa<sup>1</sup>, Kazuki Ishikawa<sup>1</sup>, Shoki Ohta<sup>1</sup>, Hajime Murai<sup>1</sup>

## Introduction

Propp has demonstrated that the plots of stories in a specific genre are combinations of finite patterns [1]. Accordingly, there have been studies that have analyzed story plots [2, 3]. These studies show the plot structure to be unique to a specific genre. However, a comparison of similar genres has not been sufficiently performed. In this research, in order to analyze the patterns that found in typical horror and detective stories, both include mystery in the story structure, we have compared the structure patterns of “Case closed” [4] and “Thriller restaurant” [5] via the use of 4-gram.

## Target Contents and Method Used

The target content was collected based on sales. As a sample of a detective story, volumes 1 to 45 of “Case Closed” were chosen. As a sample of a horror story, 15 books of “Thriller Restaurant” series was chosen.

For both the genres, the story structure was extracted via the reading of target contents. First, the story was divided into quests based on the protagonist’s goal or purpose [6]. A total of 134 quests of detective stories and 167 quests of horror stories were collected. Each quest was divided into a scene, and each scene was tagged based on the main function of the scene in question. None of the scenes had two or more tags. Each function of a scene was tagged as a plot element. The number of tags used was 29. Table 1 presents the major plot elements along with the description of their functions. Table 2 represents the number of quests, scenes, and scenes per quest for both the detective and horror genres. A 4-gram analysis of the plot sequences was conducted.

## Results

Table 3 presents the five patterns most frequently used in the detective genre. Table 4 presents the eight patterns most frequently used in the horror genre.

These results have been taken into consideration. Figure 1 presents a typical pattern of a detective story. Figure 2 presents a typical pattern of a horror story.

The tag specific to the detective genre is “Search.” The reason for this could be the intention of characters. In detective stories, the detective is motivated to close the case. However, in horror stories, the characters are not necessarily motivated to discover the truth behind the horrific experiences. The tags specific to the horror genre are “Doubt” and

---

<sup>1</sup> Future University Hakodate

“Leaving.” “Doubt” is specific because “Doubt” in a detective story could be included in “Search” or “Discovery” contexts. “Leaving” is specific to a horror story because a horror story directly presents the scenes wherein the characters are “Leaving,” such as die or disappear. On the contrary, a detective story does not present the scene in which the victims are killed because the murder is revealed afterwards. These characteristics show that both detective as well as horror genres include mystery in the story structure. While the detective genre aims to reveal the truth, the horror presents the reactions of characters to mysterious experiences.

**Table 1. Categories of plot elements**

Arrival	Encounter with the protagonist, includes elements such as birth and revival.
Leaving	Leaving from the story, includes permanent leaving elements such as death.
Losing travel route	A character cannot move, includes elements such as losing transportation facilities, detention, kidnapping, and arrest.
Search	Effort made to obtain information, includes elements such as exploration, survey, and research.
Discovery	Disclosure of some information or hidden truth.
Doubt	One of the character notices a suspicious point and doubts.
External information	External information presentation for audiences, includes elements such as prologue, epilogue, and explanation of the world of the story.
Intention, request	It includes the scenes related to the characters’ decision making, such as wishing, requesting, persuasion, and invitation.
Completion of request	A scene that mainly describes the completion of a request.
Interference	It not only includes explicit interferences, but also a few acts that intentionally make the other person uncomfortable.
Everyday	Scenes that describe only the ordinary everyday life.

**Table 2. Quests and scenes of each of the genres**

	Quests	Scenes	Scenes/Quests
Detective	134	1271	9.5
Horror	167	1487	8.9

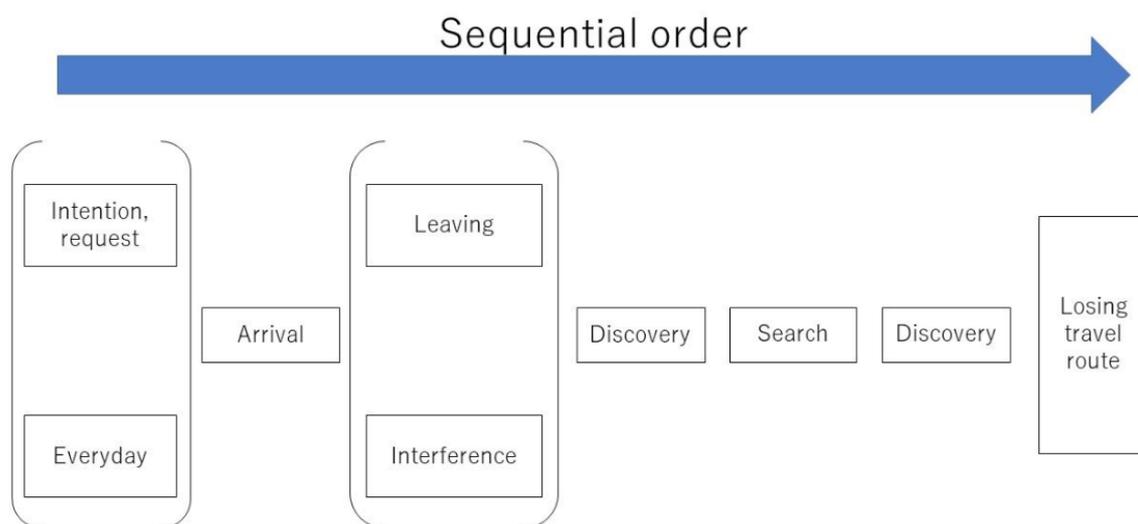
**Table 3. Frequently used patterns in the detective genre**

Rank	Elements	Appearances
1	Arrival→Discovery→Search→Discovery	59
2	Intention, Request→Discovery→Search→Discovery	50
3	Intention, Request→Arrival→Search→Discovery	46
4	Discovery→Search→Discovery→Losing Travel Route	44
5	Arrival→Search→Discovery→Losing Travel Route	41

**Table 4. Frequently used patterns in horror genre**

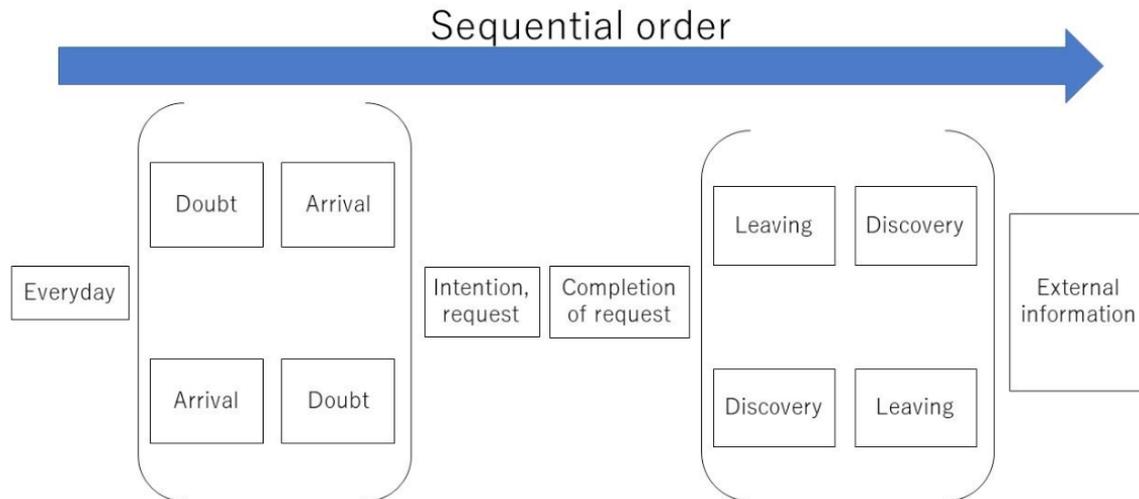
Rank	Elements	Appearances
1	Everyday→Arrival→Leaving→Discovery	12
2	Arrival→Intention, Request→Discovery→Leaving	11
2	Everyday→Discovery→Leaving→Discovery	11
4	Arrival→Intention, Request→Completion of Request→Discovery	10
4	Doubt→Arrival→Intention, Request→Discovery	10
4	Intention, Request→Arrival→Intention, Request→Leaving	10
4	Arrival→Intention, Request→Leaving→Discovery	10
4	Arrival→Doubt→Arrival→Leaving	10

## Detective



**Figure 1. Typical pattern of a detective**

## Horror



**Figure 2. Typical pattern of a horror story**

### Conclusion

In order to compare the story structures of genres include mystery in the story structure, “Case Closed” and “Thriller Restaurant” have been analyzed and symbolized. A 4-gram analysis was applied to the symbolized story structures. Then the patterns that are frequently used in both the genres were extracted based on 4-gram result. In addition, the structural differences between the two genres were considered. As a result, it is suggested that detective genre aims to solve mystery while horror genre focuses to character’s reaction.

### References

- [1]. Propp, V., *Morphology of the Folk Tale*. U of Texas P, USA, 1968.
- [2]. Murai, H., "Plot analysis for describing punch line functions in Shinichi Hoshi's microfiction." 2014 Workshop on Computational Models of Narrative. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2014.
- [3]. Toyosawa, S., & Murai, H., “Narrative structure analysis punchlines of SF genre within the flash fiction of Shinichi Hoshi.” The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, p.3L3-OS-22a-03, 2019. (In Japanese).
- [4]. Aoyama, G., *Case Closed*. Shogakukan, Japan, 1994.
- [5]. Matsutani, M., *Thriller Restaurant*. Doshinsha, Japan, 1996.
- [6]. Nakamura, S., & Murai, H., “Proposal of a method for analyzing story structure of role-playing games focusing on quests structure.” *Computer and Humanities Symposium*, vol.2020, pp.149-156, 2020. (In Japanese).

# Using Moodle as a Multi-Modal Tool for Ainu Language Education

Matthew Cotter<sup>1</sup>, Takayuki Okazaki<sup>2</sup>, Jennifer Teeter<sup>3</sup>

## Introduction

"It takes only one generation to lose a language and at least three generations to restore that language." — Tariana Turia, 2011

Launched in March 2021, the online Te Ataarangi Ainu Language class is the first of its kind. The class attracts 20-30 participants a week from locations inside and outside of Japan and was developed to bring Ainu language learners and teachers interested in strengthening the Ainu language in the face of its status as a highly critically endangered language by UNESCO (2009). Through weekly Zoom lessons broadcast live from Nibutani, Hokkaido, classes utilized a tried and proven method, the 'silent way' language learning pedagogy and take advantage of an online open-access learning management system (LMS), Moodle, with the goal of enhancing the language learning experience. The multi-modal functionality of Moodle serves as both an opportunity for self-access learning and revision of the Zoom classes and also data storage for future availability.

## Approach

The silent way, first proposed by Caleb Gattegno (1963) focuses on utilizing silence, gestures and props to elicit attention and active participation from language learners. The method has also been used by several indigenous groups notably Maori in New Zealand, who adapted it to their cultural and historical contexts. Developed by a Māori author and educator, Kāterina Te Heikōkō Mataira and a community leader, Ngoingoi Pewhairangi in 1970s, practitioners have been active in providing training in this highly successful technique to people working in endangered language revitalization (Okazaki, 2015), including Ainu teachers and participants in the aforementioned Ainu language class.

The online Ainu language class itself involves one 90-minute class per week via the online video conferencing tool, Zoom. Self-access revision and study materials are provided through employment of Moodle, a Learning Management System (LMS). Participants include Ainu and non-Ainu of multiple nationalities, and range from Ainu enrolled in a three-year apprentice program in Biratori, especially in Nibutani, where grassroots Ainu language revitalization efforts are made, shopkeepers, primary school

---

<sup>1</sup> Hokusei Gakuen University Junior College

<sup>2</sup> Kinki University

<sup>3</sup> Kyoto Seika University

teachers, university professors, university students, museum staff at Upopoy: the Symbolic Space for Ethnic Harmony (Shiraoi, Hokkaido) and people dedicated to Ainu language revitalization.

Figure 1 shows how the central teacher uses the Cuisenaire rods and props to first teach the class content on Zoom for each session. Up to five more co-teachers guide practice and revision of that content in Zoom breakout rooms. The main tools used in the online classes are Cuisenaire rods, commonly used for the silent way, and various props to signify who is talking to who during the scripted conversations, something that needs to be established when within the live video Zoom context.

Figure 2 shows the Moodle course page developed for the project. Moodle was chosen as the preferred LMS by Ainu language teachers and researchers due to the range of functions, or multi-modal features, that were required to suit both the teacher and student needs and also user level. A course was created and after registering both teachers and students, a quick explanation of how to use the course was performed during one of the Zoom sessions. Main functions used on Moodle are 1) the video upload activity for data storage and to enable students to access and review the online classes, 2) downloadable pdfs of class content such as new words and phrases learned during the classes 3) forum and video assessment module for submitting desired content, and 4) quizzes to test and assess learned content.



**Figure 1: Central teacher uses Cuisenaire rods and props to teach the class content via Zoom**

4 **2021pa 4cup 20to inanike e=e rusuy? tanpe kina ka somo ne**

2021.4.20 iporse

7 iporse(sisam itak tura)2021.4.20

iporse quiz: 2021pa 4cup 20to

**Restricted** Available from 20 April 2021, 11:00 PM

21.4.20cinumkekampi

2021.4.20ausarayetumpu

tepakno a=ki p opitta

poro tumpul (inanike e=e rusuy)

**Figure 2: Moodle course page developed for revision of content and data storage**

## Conclusion

In this presentation, the background and history into the need for the course will be summarized. The method and style of the online classes will also be described and how Moodle was used to facilitate self-access and revision of content. This project is part of the Grants-in-Aid for Scientific Research (KAKENHI) project ‘Improving Awareness Understanding of Ainu via Online Resources’ number 20K01208.

## Reference

- Gattegno, C. (1963). Teaching Foreign Languages in Schools: The Silent Way (1st ed.). Reading, UK: Educational Explorers. Retrieved from [https://issuu.com/eswi/docs/gattegno\\_-\\_teaching\\_foreign\\_languages\\_in\\_schools\\_t](https://issuu.com/eswi/docs/gattegno_-_teaching_foreign_languages_in_schools_t)
- Okazaki, T. (2015). Te Ataaragi to Maorigofukko [Te Ataarangi and Maori language Revitalization]. Konton 12. pp.48-65
- Turia, T. (2011). Maori Development Organisation, Speech at 13th Annual Provider Awards 2011 Pohutu Cultural Theatre, Whakarewarewa, Rotorua
- UNESCO (2009). UNESCO Interactive Atlas of the World’s Languages in Danger, Retrieved from <http://www.unesco.org/culture/ich/index.php?pg=00206> 1

# **An Attempt at Creating Integrated Retrieval for Chinese Excavated Materials: An Implementation of a Search Function across Interpretations of Ancient Characters**

**Shumpei Katakura<sup>1</sup>**

## **Background**

This paper describes a research on development of an integrated search function for comprehensively finding a phrase from ancient Chinese text data in which each ancient character has multilayered interpretation data.

Many excavated materials from the Warring States period (5th to 3rd centuries B.C.) have been rediscovered in mainland China since the 1950s. These materials are essential to studies on ancient China. For research purposes, we attempt to digitalize information about them as early as possible. However, it is difficult to interpret these ancient characters because many usages of them differ from those of Hanzi(漢字), which we use now. Excavated materials are found and reported frequently, so arguments on interpreting many difficult ancient characters have been accumulated without defining correct answers.

When information about newly excavated materials is released, first, we begin by interpreting the characters and sentences on them. Then, we refer to other excavated materials since we cannot interpret correctly until we examine enormous interpretations on other materials, such as finding characters that have similar shape and sentences, which have similar construction. To improve the efficiency of interpreting through digitalizing, we create a system that contains many interpretations on the excavated materials and can easily find them.

## **Method**

In this search function, each character on the excavated materials has multilayered interpretation data called "Liding"(隸定) and "Shidu"(稊讀). "Liding" is the interpretations pertaining the shape of characters: here, components of ancient characters are modified to those of Hanzi. "Shidu" is the interpretations on the meaning of characters: the use of an ancient character is examined as corresponds to that of Hanzi (Figure 1). It is helpful to create this system because we can find characters and sentences on excavated materials in terms of both shape and meaning.

---

<sup>1</sup> The University of Tokyo

**Figure 1: An example of “Liding” and “Shidu”.**

For example, suppose that there is a Liding sentence on an excavated material as "... 又隈迓...", and there are various Shidu interpretations about 「隈」 and 「迓」. (In fact, this sentence does not exist. We created this character string for explanatory convenience.) Table 1 shows the interpretations on each character with Liding and Shidu by three researchers (A, B, and C).

**Table 1: The interpretation data of "...又隈迓...".**

	Liding	Shidu A	Shidu B	Shidu C
...	...	...	...	...
①	又	有	有	有
②	隈	魏	悞	威
③	迓	逆	苗	朝
...	...	...	...	...

Here, our system finds this sentence using any search query as below.

- (1): To input "又隈" ... Liding of ① and Liding of ② come up.
  - (2): To input "威苗" ... Shidu C of ② and Shidu B of ③ come up.
  - (3): To input "又魏" ... Liding of ① and Shidu A of ② come up.
  - (4): To input "有\*逆" ... Shidu A, B, and C of ① and Shidu A of ③ come up.
- The function "\*" is the wildcard.

**Table 2: The diagram of search example.**

	Liding	Shidu A	Shidu B	Shidu C
...	...	...	...	...
①	又	有	有	有
②	隈	魏	悞	威
③	迓	逆	苗	朝
...	...	...	...	...

Depending on our system, we can access various interpretations for search, which include searching only from Liding data like (1), only from Shidu data like (2), across Liding and Shidu data like (3), across arbitrary characters like (4) (Table 2). To search for any phrase, using the functions (1), (2), and (3) are effective. To search for any co-occurrence

character, the function (4) is effective. By registering as many data as possible, we can show more information about the characters and sentences.

Here is a case where this system works effectively. In one of the excavated materials, "Baoshang Chu Bamboo-slips(包山楚簡)", there are two Liding phrases as "魯易" (slip number 1) and "[𠄎𠄎旅]易" (slip number 4). The shapes of these ancient characters are critically different, however, many interpretations suggest they are both the same person, "魯陽", who appears in some Chinese classics, such as "Guoyu(國語)", therefore we can examine "魯陽" as Shidu of both "魯易" and "[𠄎𠄎旅]易"[2]. To adopt our system, which enables to search from Shidu data, you can simultaneously find "魯易" in slip number 1 and "[𠄎𠄎旅]易" in slip number 4 by inputting "魯陽".

Further, we create original text dataset with metadata (e.g., split number, IDS information) for search convenience (Table 3).

**Table 3: An example of original text dataset of excavated materials.**

	A	B	C	D	E	F	G	H
1	type	title	slip number	orders of appearance	Liding	IDS	Shidu(Jiagu)	note
354	文書	集筭言	15		32 登	𠄎𠄎豆		
355	文書	集筭言	15		33 壘	𠄎𠄎土		
356	文書	集筭言	15		34 而			
357	文書	集筭言	15		35 無			
358	文書	集筭言	15		36 古		故	

## Conclusion

Many text data of excavated materials on some websites are now described only as Liding data; even if Shidu data are written together, only one Shidu interpretation judged by the publisher as most suitable is encompassed. Upon operation, our system puts together many Shidu interpretations comprehensively, which will lead to digitalizing text data of excavated materials. We plan to correlate both text and image data and incorporate this search system into our digital archive on development.

## Reference

- [1]. 李学勤, 清華大学出土文献研究与保護中心. 清華大学藏戰国竹簡 8, 中西書局, 2018.
- [2]. 劉信芳. 包山楚簡解詁, 藝文印書館, 2003.

# Collecting Canons: Comparing Guodian and Mawangdui *Laozi* Texts with the Dead Sea Scrolls

Janelle Peters

## Project Description

This project allows students to interact with the manuscripts of the *Laozi* and the Dead Sea Scrolls. It uses extant scholarly collections of the Guodian Bamboo Texts, the Mawangdui Silk Texts, and the Dead Sea Scrolls. It gives students a set of maps, timelines, themes, and networks of authority. This allows students to compare the contemporary *Laozi* and Dead Sea Scrolls through the lens of their respective (and largely unconnected) contexts, the lens of receiving communities, and the lens of digital humanists.

## Corpora

The *Laozi* texts were composed in roughly the same period (4<sup>th</sup> c. – 2<sup>nd</sup> c. BCE) as the Dead Sea Scrolls (3<sup>rd</sup> c. BCE – 1<sup>st</sup> c. CE). Both sets of texts exhibit group blending and change. For the *Laozi* texts, one finds Confucianism and Daoism. For Qumran, scholars still debate the exact nature of the insular and yet widespread nature of the Judaism of the group(s) that collected the scrolls. All of these texts have the advantage of having intentionally collected and placed manuscripts; they were not sifted out of the trash, leaving scholars to puzzle about their continued significance and the reasons for discarding them. They stretch over a small number of centuries, allowing scholars to find subtle changes such as the addition of ideals of emulating water or feminine ways between the Guodian and Mawangdui manuscripts despite an overall trend toward textual conservatism. [1]

## Pedagogical History

This project comes out of several semesters teaching core classes to non-majors at the lower-division and upper-division undergraduate levels at Loyola Marymount University in Los Angeles, California. Students often encounter texts for the first time, and they do not have a consistent training in digital humanities methodology. They need classes that will allow them to engage in the details of the text in a meaningful way while also not presupposing any previous knowledge on their part. These core classes have two challenges: 1) familiarizing students with voluminous textual traditions, 2) furthering students in knowledge of scholarly techniques.

For the first challenge, there is the fact that the early *Laozi* manuscripts and the Dead Sea Scrolls have a large set of manuscripts. While hardly “Big Data,” the

manuscripts nonetheless have a scale that is too large for undergraduate students to read in their entirety in a non-major semester course. Moreover, there is evidence that the *Laozi* and the sacred texts found at Qumran were reordered in ways that are not found in later canonical traditions. Allowing students to manipulate the manuscript content allows them to interact with the texts in a process more aligned to what transpired in antiquity and accelerates their exposure to textual themes through their own analyses.

For the second challenge, there is a classroom context in which many of the undergraduate students apply advanced scholarly techniques for the first time. The compendious site <http://www.bamboosilk.org/> is in the Chinese language, which is not a language most of the students know. There are some sites that allow students to see the multiple translations of later traditions (<https://pages.ucsd.edu/~dkjordan/chin/LaoJuang/DDJTenTranslations.html>), but undergraduate, non-major students need guided projects to reorder texts and compare themes. Such work generates helpful insight into the imagination and enduring humanistic value of the texts and also the issues involved in historiography, translation, and the digital humanities.

## Method

Students will use R and Palladio to create word clouds, timelines, networks (e.g., Teacher of Righteousness), and maps. Digital humanities problems the students will have to consider will range from data cleaning (including grammatical number) to lemmatization (including continuous script concerns) to named-entry recognition. A specific contribution the students can make is to include Dead Sea Scrolls texts with community beliefs and Dead Sea Scrolls texts now considered part of the Catholic or Orthodox canon fully in the analysis, such as the book of Tobit. [2] Scholarly analyses customarily bracket out these texts due to their lack of inclusion in the rabbinic Jewish canon and also the Protestant canon, though they quite helpfully have been taking the form of Palladio-ready lists and Excel spreadsheets for decades. [3] Nonetheless, like scholarly discourse on the development of the *Laozi*, scholars do not wholly agree upon which texts were considered to have canonical form at Qumran, but they are certain that these texts were considered important enough to be preserved among scriptural texts and community documents in a setting oriented toward a belief in divine interaction. [4] None of these sites are like that of Nag Hammadi, and their assemblages therefore all have more intrinsic coherence as they were intended to be collected and preserved at their locations. Moreover, it is possible that the texts many Jews during the Second Temple Period considered canonical were not exactly identical to the later Hebrew Bible. [5]

### Desired Outcome

Learning outcomes should yield an appreciation for diverse “canonization” or standardization processes and textual fluidity, a consideration of the work of scholars in preparing editions and translations, and familiarity with exploring aspects of lists through use of R and Palladio. Students do not have to make a major breakthrough with this project. They are learning about the situated texts by concentrating on the details of the sites and the texts and asking their own questions within the framework of R and Palladio.

### Reference

- [1]. Edward L. Shaughnessy, “A First Reading of the Shanghai Museum Bamboo-Strip Manuscript of the Zhou Yi,” *Early China* 30 (2005): 1-24; Edward L. Shaughnessy, *Rewriting Early Chinese Texts* (SUNY 2006).
- [2]. Joseph Fitzmyer, “The Aramaic and Hebrew Fragments of Tobit from Qumran Cave 4,” *Catholic Biblical Quarterly* 57 (1995): 655-675.
- [3]. Emanuel Tov, *Revised Lists of the Texts from the Judaean Desert* (Brill 2009), chapter 3.
- [4]. Guolong Lai, *Excavating the Afterlife: The Archeology of Early Chinese Religion* (Washington 2015), chapter 5.
- [5]. Hindy Najman, “The Vitality of Scripture Within and Beyond the ‘Canon,’” *Journal for the Study of Judaism* 43 (2012): 497-518; Tim Whitmarsh, *Dirty Love: The Genealogy of the Ancient Greek Novel* (Oxford 2018), 103.

# Development of Database for Japanese Conversation Patterns: an observation from noun phrases ending with focus particle "*mo* (also)"

Mika Ebara<sup>1</sup>, Hilofumi Yamamoto<sup>1</sup>

## Introduction

The purpose of the present study is to propose the development of a database for Japanese conversational patterns and rules using a recorded reconstruction discourse dataset. To explain the necessity of the development, we will show an example of the analysis of noun phrases with a focus particle (*mo*: also). Research studies have been very actively studying Japanese focus particles or binding particles (Aoyagi, 2008; Numata, 2009). Although many studies have been conducted on the functions of particles in sentences, focus particles are difficult to deal with in the original formats of recorded linguistic texts since the particles indicate the focus element in a sentence. Since not only particles but also content words are often omitted in Japanese conversation, we will reconstruct full sentence to start the analyses.

## Application Example

The present paper will analyze the particle “*mo*” occurring after a noun phrase in the final position of an utterance in a casual conversation. We will show an example from the corpus data of the Gen-Nichi-Ken Corpus of Workplace Conversation (CWPC). The search conditions were as follows: a string consisting of noun phrase + binding particle and one word immediately before the end of the utterance unit. Eighteen cases were matched in total. Table 1 indicates three extracted cases where the utterances ended with *mo*, which are #3, #4, and #11.

Since each utterance ends before the appearance of a noun phrase with *mo*, the results allow us to estimate two possibilities of sentence construction, that is, a one-word sentence and a right dislocational sentence. The one-word sentence in the context indicates an answer, confirmation, or question for the question, topic presentation, or time/place setting (Onoe, 1998). However, the previous context for each of #3, #4, and #11 does not contain question equivalent sentences, nor does “English”, “23”, or “the first time” represent any question, topic, time or place, respectively. None of these elements can be regarded as features of a one-word sentence.

---

<sup>1</sup> Tokyo Institute of Technology

**Table 1: The ending fragment of each utterance unit is accompanied with noun phrase + *mo*; texts taken from Gen-Nichi-Ken Corpus of Workplace Conversation (CWPC)**

ID	Japanese	English
3.	PN, Eigo mada dekinai, osietekure tte	PN (said) I cannot do English yet, teach me
	Honto	really
	Kotti de osieta no mo aru	there is a case that I taught
	Eigo mo	<KEY> English too
	Un	yeah
4.	23 desyo, de, 25 desyo	23(rd), and 25(th)
	25 wa ne, kekkoo minna deteru yo	25(th), everyone is out
	23 mo	<KEY> 23(rd), too
	PN san mo ne	PN, too
11.	Kore doo datta kke	<Q> how was this?
	Konna datta ke	<Q> like this?
	A, sonna desita	oh, like that
	Saisyo mo	<KEY> the first time too
	Saisyo no yatu desuka	the first one

We will examine another possibility: the right dislocational sentence. The right dislocation can be analyzed as follows: there are two sentences where any predictable part is omitted; and any element in a sentence is dislocated to the right or left of the original position (Watanuki, 2006). If some parts in a sentence are omitted phonologically, all cases can be recovered. In addition, if the omitted part is presumed, it can be recovered in the same way according to what they uttered previously. We will reconstruct sentences with word-for-word glosses, such as in (1), which is the extracted utterances from #3. The missing particles and words need to be complemented, as in (1b).

- (1) a. PN(personal name), Eigo mada deki-nai, osiete-kure tte  
 PN English yet can-NEG teach (for me) that  
 SUBJ OBJ ADV AUX VERB QUOT
- b. PN wa Eigo ga mada deki-nai, osiete-kure tte itta  
 PN-TOP English-OBJ ADV AUX VERB QUOT VERB

We reconstructed the sentences in three cases and marked if they were acceptable after the reconstruction as shown in Table 2. As a result, the sentences in #3 and

#4 were not acceptable. A sentence that includes a noun phrase with *mo* may not always be a dislocation phenomenon. The noun phrase with *mo* in #11 can be regarded as naturally embedded in the previous sentence. We can analyze them using a syntactic approach. #11 has a syntactic property which is a right dislocation but #3 and #4 cannot be explained within the syntactic analysis. Hence, the sentence reconstruction we attempted is one of the indispensable processes for describing underlying grammatical rules.

**Table 2: Data reconstructed to full sentences and their acceptances. \* indicates unacceptable data; ✓ indicates acceptable data; texts taken from Gen-Nichi-Ken Corpus of Workplace Conversation (CWPC)**

ID	Japanese	Full sentence and acceptance
3.	PN, Eigo mada dekinai, osietekure tte	PN wa Eigo ga mada dekinai, osietekure tte itta
	Honto	Honto ni
	Kotti de osieta no mo aru	Kare ni kotti de osieta no mo aru
	Eigo mo	* <u>Kare ni Eigo mo osieta</u>
	Un	Un
4.	23 desyo, de, 25 desyo	23 desyo, de, 25 desyo
	25 wa ne, kekkoo minna deteru yo	25 wa ne, minna wa kekkoo soto ni deteru yo
	23 mo	* <u>Minna wa 23 mo soto ni deteru</u>
	PN san mo ne	PN san mo soto ni deteru ne
11.	Kore doo datta kke	Kore wa doo datta kke
	Konna datta ke	Kore wa konna no datta kke
	A, sonna desita	A, kore wa sonna no desita
	Saisyō mo	✓ <u>Saisyō mo sonna no desita</u>
	Saisyō no yatu desuka	Saisyō no yatu desuka

## Conclusion

One of the three data extracted from CWPC can be analyzed as a right dislocation, whereas two data show counterexamples to previous sentence-based studies. To obtain these findings, we manually transformed the original format of the conversation corpus into one consistent searchable string and attached marks to indicate if the reconstructed sentences were acceptable in the context. We conclude that the conventional elements need to be reconstructed into a searchable unit and compiled as a database. In the future, we will use other corpora together and develop a research database that includes the description of search conditions and research questions.

## References

- [1]. Aoyagi, H. (2008). On the Morphological, Syntactic and Semantic Behavior of Toritate Particles: With an Emphasis on the Distinction between Kakari-Joshi and Fuku-Joshi. *Journal of Japanese Grammar*, 8(2): 37-53.
- [2]. Gen-Nichi-Ken Corpus of Workplace Conversation (CWPC) <https://chunagon.ninjal.ac.jp/shokuba/search> (Accessed on 2021-06-07). National Institute for Japanese Language and Linguistics. (Database from Gendai Nihongo Kenkyu Kai Ed. (2011). *Gappon Josei no Kotoba Dansei no Kotoba (Shokuba Hen)* [Combined Book: Expressions of Women Expressions of Men (Workplace)]. Tokyo: Hituzi Syobo.)
- [3]. Numata, Y. (2009). *Gendai Nihongo Toritate-shi no Kenkyu* [Studies on Modern Japanese Focus Word]. Tokyo: Hituzi Syobo.
- [4]. Onoe, K. (1998). Ichigo-Bun no Yoho: “Ima / Koko” o Hanarenai Bun no Kento no tame ni [Usage of One-Word-Sentence: Consideration of Sentence Depending on “Now / Here”]. *Tokyo Daigaku Kokugo Kenkyushitsu Sosetsu Hyaku Shunen Kinen Kokugo Kenkyu Ronshu* [100th Anniversary of Tokyo University Laboratory of Japanese Language: Collections of Japanese Language Study], 888-908. Tokyo: Kyuko Shoin.
- [5]. Watanuki, K. (2006). Nihongo no Kochibun: Sahoidobun to no Soi [Japanese Right Dislocation: Difference from Left Dislocation]. *Scientific Approaches to Language*, 5: 251-268.

# Drug-focused text summarization of coronavirus-related articles for the discovery of COVID-19 therapies

Setsuro Matsuda<sup>1</sup>

## Abstract

Since December in 2019, we humans have been exposed to the threat of COVID-19. As of May 17, 2021, about 163 million people have been infected with the novel coronavirus. Although the vaccination against this virus has already started in various countries, no effective drugs for treatment have been developed to date. What is worse, mutant variants that have a higher infection rate than the original virus emerged and are rapidly spreading all over the world. Under such circumstances, a data repository project entitled “CORD-19” has been launched to boost researches related to COVID-19. CORD-19 stands for COVID-19 Open Research Dataset and provides more than 280,000 scholarly articles about the novel coronavirus. The author downloaded the CORD-19 file released on March 8, 2021 and attempted to discover promising treatment drugs by using text mining techniques. Namely, the present work aims to repurpose or reposition existing drugs for COVID-19 therapies. The first step of this study was to extract the abstracts of the CORD-19 articles saved in the JSON format. This was done because the repository contains too many text data to be analyzed by a personal computer. After that, abstracts with “drug” or “medicine” in their texts were selected to focus on drug-related articles.

As the second step, the part-of-speech tags were added to each word of the selected abstracts with the “Tree Tagger.” The tags for proper noun: NP and NPS, which are defined in the Penn Treebank Project, enable us to identify the names of drugs or chemicals. In general, drug names tend to be lengthy and their suffix is characteristic. For example, antiviral drugs such as favipiravir and remdesivir have the suffix, -vir at the end of the name in common. Therefore, the author tried to extract proper nouns that consist of more than eight letters and end with frequently occurred three-letter suffixes such as -vir and -ine. For the purpose of double-checking, the extracted drug-like names were searched in a drug information database, “KEGG MEDICUS.” As a result, 42 proper nouns have been found to be drug names.

The third step was to identify which verbs frequently co-occur with the 42 drugs. The frequency analysis of co-occurred verbs revealed that verbs such as *convert*, *treat*, *inhibit*, *bind*, and *dock* seem to be used specifically in the context of medicine. Among the top ten verbs with a frequency higher than 60, *treat* and *inhibit* were focused since the

---

<sup>1</sup> Department of Humanities, National Institute of Technology, Matsue College, Japan

other verbs except for *convert* often appear in general contexts. The verb, *convert* was also excluded in the following analysis. This is because the verb must be part of the phrase, “angiotensin converting enzyme 2 (ACE2),” which is a primary receptor for the spike protein of the novel coronavirus (i.e. SARS-CoV-2).

In the final stage, the abstracts that include both the 42 drug names and the 2 collocational verbs were analyzed. In order to discover promising treatment drugs and obtain informative research findings, the LexRank algorithm was applied to summarizing the texts in the targeted abstracts. This algorithm is a graph-based extractive summarization and the similarity between two sentences is estimated by the cosine similarity based on bag-of-words model. Then, sentences to be included in the summary are highly ranked by the PageRank score employed in Google’s search engine. The resultant summary has demonstrated that a stereoisomer of nelfinavir, which is an antiviral drug against HIV, is an effective structure for the treatment of COVID-19 and the tyrosine kinase inhibitor, imatinib is a possible therapeutic drug for the disease. It also suggests that the compound of remdesivir and chloroquine seems to have no beneficial effect for COVID-19 patients. These results show that the methodology proposed in this study is helpful to efficiently retrieve drug information for the treatment of COVID-19 from a large-scale dataset.

# Reconstruction and Utilization of Text Data Using TEI: Case study of the *Shibusawa Eiichi Denki Shiryo*

Boyoung Kim<sup>1</sup>, Satoru Nakamura<sup>2</sup>, Yuta Hashimoto<sup>3</sup>, Naoki Kokaze<sup>4</sup>, Sayaka Inoue<sup>5</sup>, Toru Shigehara<sup>6</sup>, Kiyonori Nagasaki<sup>7</sup>

## Introduction

The *Shibusawa Eiichi denki shiryō*, biographical materials on Shibusawa Eiichi (1840-1931), who was a leading figure in the development of Japan's modern society, was published in 68 volumes between 1955 and 1971 and contains more than 38,000 primary sources. Since this material is a basic source for research on modern Japanese history, there are great expectations for its potential use when digitized. Work on digitization began in 2004 and volumes 1 to 57 of the main volume (volume 58 is excluded because of the index) can be viewed online from 2016[1]. On the other hand, the remaining 10 supplementary volumes that contain diaries (include schedule), letters, discourse, lectures, and photographs are not released yet.

## Purpose of the Study

The purpose of this study is to reconstruct the text data of the diary and schedule listed in the first and second volumes of the 10 supplementary volumes based on the Text Encoding Initiative (TEI) guidelines to provide an open access and increase usability. It was conducted for 11 months as an incentive study for Integrated Studies of Cultural and Research Resources at the National Museum of Japanese History in 2020[2]. The goals of the research are (1) to propose a versatile TEI markup method for Japanese materials, (2) to present a multifaceted research approach through various visualizations and analyses, and (3) to consider a possibility of application to archival materials. In this presentation, we will describe the research contents and results of these goals.

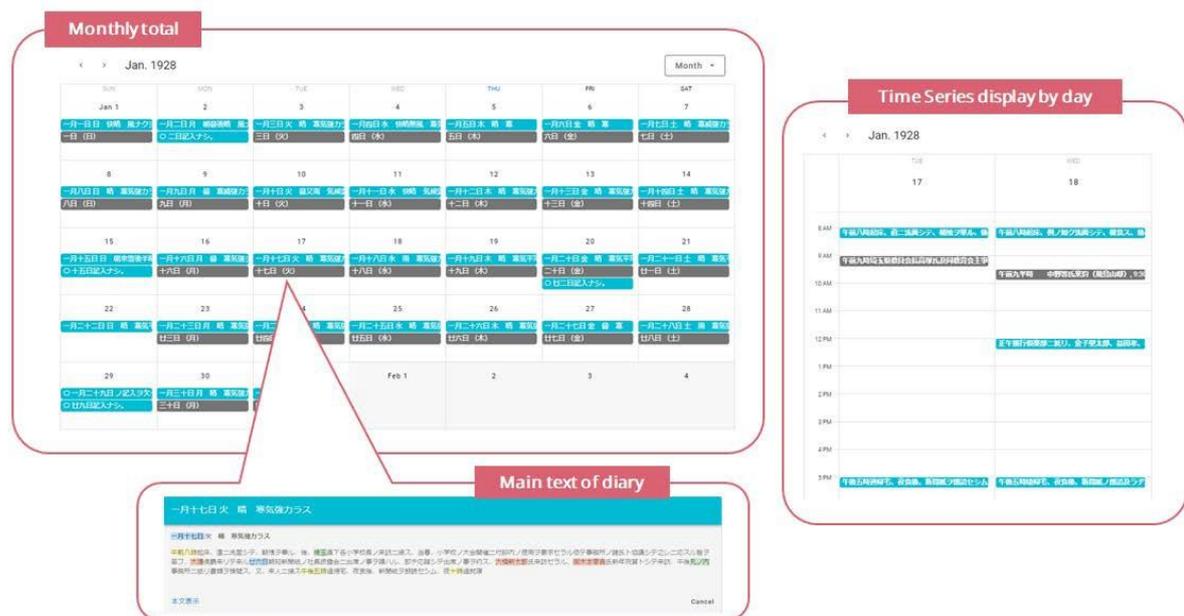
## Contents and results

First, for the goal of (1), we examined the markup method for data structuring and Japanese-specific expressions. Since the text data exceeds 18,000 characters, we decided

- 
- 1 Shibusawa Eiichi Memorial Foundation
  - 2 The University of Tokyo
  - 3 National Museum of Japanese History
  - 4 Chiba University
  - 5 Shibusawa Eiichi Memorial Foundation
  - 6 Shibusawa Eiichi Memorial Foundation
  - 7 International Institute for Digital Humanities

on the overall structure first and marked it up automatically using Python, and finally, fixed the TEI tag manually. Regarding data structuring, we marked up both the content structure of materials such as volumes, parts, and chapters and the physical structure of a book. In particular, it can be said that the feature of this research is that the structure was designed based on the group of primary sources. Moreover, we automatically tagged some types of entities such as person names and place names via a function of a morphological analysis tool, Mecab. With regard to Japanese-specific expressions with TEI markup, however, although some problems that arise, such as problems when writing Warigaki and vertical writing horizontally, have been extracted, the research didn't reach the stage of showing the concrete answer. It should be studied further.

Next, for the goal of (2), we worked on (a) full-text display and (b) visualization with tags for unique expressions, by using not only TEI but also the International Image Interoperability Framework (IIIF) and the Resource Description Framework (RDF). For (a) full-text display, we developed a viewer that displays the contents (table of contents) using the hierarchical structure of TEI-structured text data. In addition, an environment in which images and text can be viewed simultaneously has been developed by associating IIIF images with the text using TEI facsimile tags. (b)As for the visualization with tags for unique expressions, we used "date and time" and "person and place name" information for visualization. In the visualization using "date and time" information, we visualized the total number and distribution of data by year, and visualized them in calendar format, as shown in Figure 1.



**Figure 1: Visualization of diaries in calendar format**

In addition, for visualization using "person and place name" information, we described information about people and places using RDF and associated it with an

external knowledge base such as DBpedia Japanese to obtain thumbnail images, descriptions, and location information. By utilizing this extended information, the functions to visualize a network of people and to present items that provide evidence of the relations were developed as shown in Figure 2. Furthermore, we also provide a search function based on place names on a map using location information.

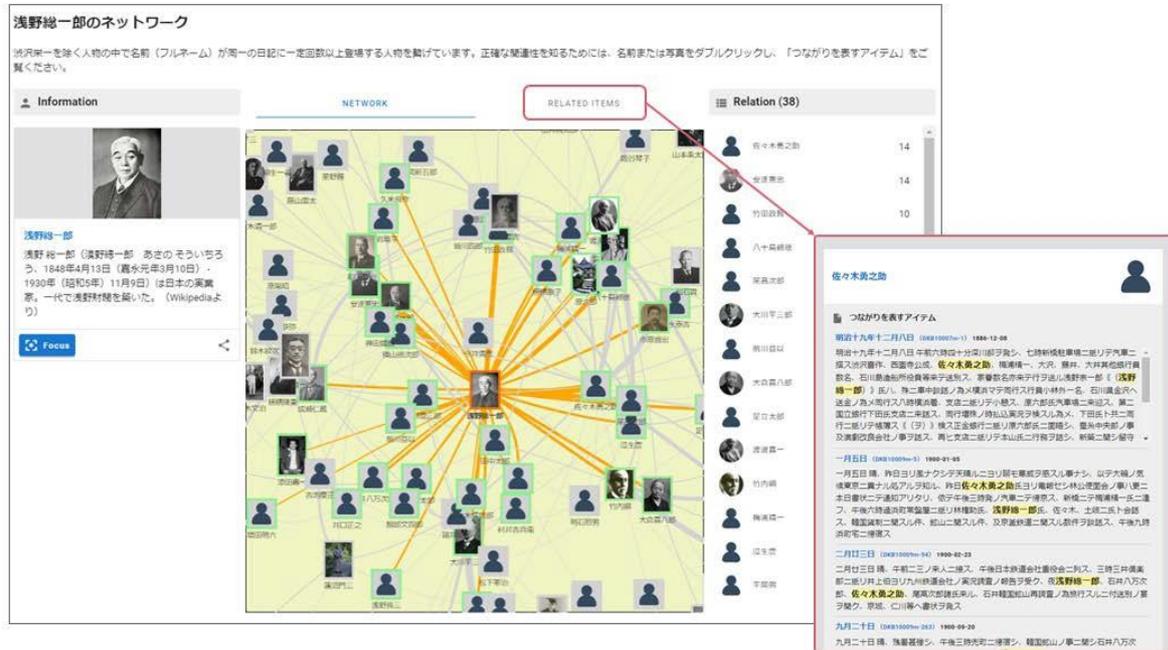


Figure 2: People network and items that provide evidence of the relations

## Conclusion

Here are three outcomes of this research. The first is the release of the full-text data which have not been published online until now. The TEI file can be obtained from the "Shibusawa Eiichi Diary"[3] built for public use. The second is to show a method that takes into account the characteristics of the text data in the TEI markup. This approach can provide a reference to archival institutions to improve the utilization of their holdings. The third is the construction and release of "Shibusawa Eiichi Diary" as a research support tool. Visualization on this site using calendars, graphs, network diagrams, and maps made it easy to analyze the text data from various points of view such as the yearly distribution of materials, the lifestyle patterns of Shibusawa Eiichi, and the connections between related people.

## Reference

- [1]. デジタル版『渋沢栄一伝記資料』, <https://eiichi.shibusawa.or.jp/denkishiryu/digital/main/>, Accessed on 2021-07-07
- [2]. 2020 年度国立歴史民俗博物館総合資料学奨励研究 「TEI を用いた『渋沢栄一伝記資料』テキストデータの再構築と活用」

- [3]. Shibusawa Eiichi Diary, <https://shibusawa-dlab.github.io/app1/>, Accessed on 2021-07-07

## Development of a support system for extracting mentioned bibliographical data from the *Encyclopédie* entries

Satoru Nakamura<sup>1</sup>, Ayano Kokaze<sup>2</sup>, Yoshiho Iida<sup>3</sup>, Naoki Kokaze<sup>4</sup>, Tatsuo Hemmi<sup>5</sup>

### Introduction

This paper presents the system, with its significance, we have developed to support the Society for the Study of the *Encyclopédie* and the Enlightenment which is working on a project (1) to identify the works mentioned in the text of each entry of the *Encyclopédie*, (2) to list them up, and finally (3) to complete their bibliographical data. The *Encyclopédie* is a large-scale encyclopedia edited by the French Enlightenment philosophers, *i.e.* Denis Diderot and Jean Le Rond d'Alembert and other *Encyclopédistes*, for more than 20 years from 1751 to 1772. It consists of 28 volumes, of which 17 contained the text and 11 illustrations, followed by a supplementary volume and an index. The total number of entries is more than 70,000. The text of the *Encyclopédie* is written by, while referring to and often quoting from, a variety of prior texts, *e.g.* Bibles, dictionaries, books, and journal articles.

Our contributions and further implications consist in the following points. (1) From the viewpoint of the *Encyclopédie* studies, ours will provide a complete list of works that the *Encyclopédistes* referred to in writing the entries and that the *Encyclopédie* as a whole is based on, which the previous research and critical editions of the *Encyclopédie* did not mainly focus on [Schwab 1971–1972; ARTFL; ENCCRE]. (2) The results of this study, when completed, will have the following historical and DH implications. This study contributes to creating citation networks for the philosophers of the Age of Enlightenment by aggregating data on the bibliographical information referred to in the *Encyclopédie* entries. Identifying the formation of knowledge through citation networks has been the subject of DH, as in contributions dealing with intertextuality in the field of classics [Murai et al. 2007; Romanello 2016]. As for historiography, there is argument from the gender perspective, *e.g.* [Kawashima 2005], that academia remained male-dominated in the 18th century, despite many women in the salons: this gender bias is a common problem with citation networks of the *Encyclopédie*. Although this issue is a historical discussion of ideas in the 18th century, it is also an issue with the actuality that connects to contemporary academia: in recent years, through citation analysis, some DH scholars have

---

<sup>1</sup> The University of Tokyo Historiographical Institute

<sup>2</sup> Ochanomizu University

<sup>3</sup> Aoyama Gakuin University

<sup>4</sup> Chiba University

<sup>5</sup> Niigata University

criticized the fact that most of the highly regarded central research in DH research and education has been conducted by a few scholars from the viewpoints of language, race, geography, and gender [Fiormonte, 2015; Earhart et al. 2020]. Finally, (3) in a technical context, this project is one practice of citizen-participatory research in the humanities. Examples include the citizen-participatory transcription project “Minna de Honkoku” [Hashimoto 2019] and the “SAT Daizōkyō Text Database” [Nagasaki 2017]. In addition, there are researches for data enrichment with Named Entity Recognition [Yoshiga 2021] and tools such as Recogito [Recogito] to support markup. While referring to their methodologies, this paper presents the system to support this project.

### Process to extract authority information

Figure 1 shows an example of an entry of *L'Encyclopédie*. Nomenclature (Entry name), Auteur mentionné (Mentioned author(s)), Titre mentionné (Name of the work mentioned), and so on, are extracted from the text. In addition, normalized data of spelling inconsistencies in the text is created to aggregate authority information. For example, Tournefort's *Inst. rei herb.* mentioned in the text is normalized to Joseph Pitton de Tournefort's *Institutiones rei herbariae* (1700). Previously, the project has managed the information described above with MS Excel. As a result, there were difficulties in collaborative work, increased input errors (typos and omissions due to manual input), and difficulty getting new entrants due to knowledge and language issues.

**BAAL ou BEL**, (*Hist. anc.*) nom qui signifie *seigneur* en langue Babylonienne, & que les Assyriens donnerent à Nemrod, lorsqu'après sa mort ils l'adorerent comme un Dieu. *Baal* étoit le dieu de quelques peuples du pays du Chanaan. Les Grecs disent que c'étoit Mars, & d'autres que c'étoit ou Saturne ou le Soleil. L'historien **Joseph** appelle le dieu des Phéniciens *Baal* ou *Bel*, dont **Virgile** parle dans l'*Enéide* comme d'un roi de Tyr :

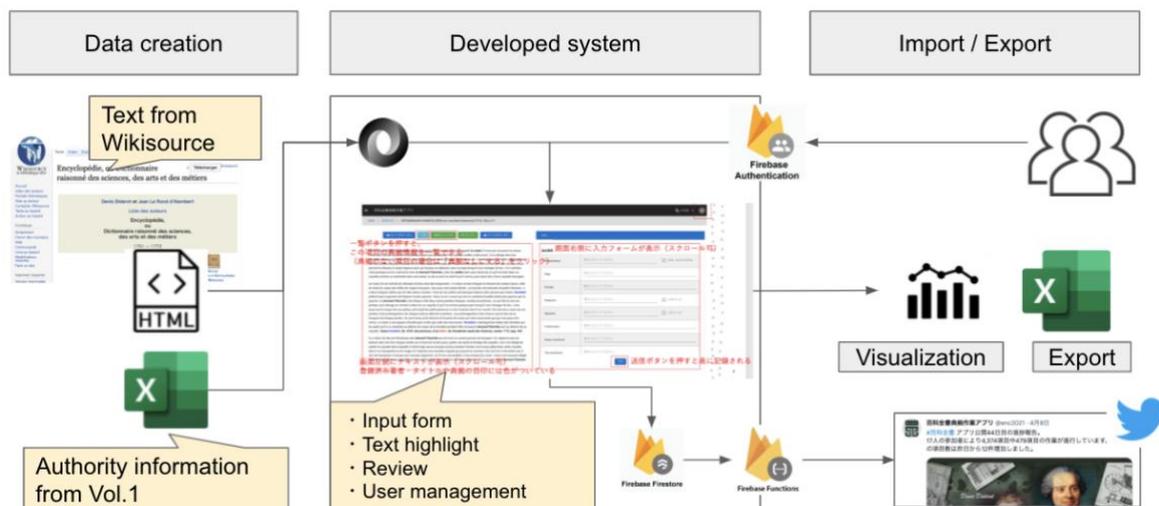
*Implevitque r* Mentioned author(s) *m Belus, & omnes A Belo s* Mentioned author(s) Name of the work mentioned

Godwin, fondé sur la ressemblance des noms, croit que le *Baal* des Phéniciens est le même que *Moloch* : le premier signifie *seigneur*, & le second, *prince* ou *roi*. Cependant d'autres pensent que ces peuples adoroient Saturne sous le nom de *Moloch*, & Jupiter sous celui de *Baal* : car ils appelloient ce dernier dieu, *Baal semen*, le *seigneur du ciel*. Quoi qu'il en soit de ces différentes opinions, le culte de *Baal* se répandit chez les Juifs, & fut porté à Carthage par les Tyriens ses fondateurs. On lui sacrifioit des victimes humaines, & des enfans, en mémoire de ce que se trouvant engagé dans une guerre dangereuse, il para son fils des ornemens royaux, & l'immola sur un autel qu'il avoit dressé lui-même. **Jérémie** reproche aux Juifs qu'ils brûloient leurs enfans en holocauste devant l'autel de *Baal* ; & dans un autre endroit, que dans la vallée d'Ennon ils faisoient passer leurs enfans par le feu en l'honneur de *Moloch*. Les Rabbins pour diminuer l'horreur de cette idolatrie, s'en sont tenus à cette seconde cérémonie. *Non comburebant illos*, disent-ils de leurs ancêtres, Name of the work mentioned *los per ignem*. Mais si dans le culte de *Baal* il n'en coûtoit pas toujourns la vie à moins étoient souvent teints du sang de ses propres prêtres, comme il paroît par le rameux sacm. Elle les défia. *Incidabant se juxta ritum suum cultris & lanceolis, donec profunderentur sanguine. Lib. III. Reg. Voyez BELUS.* (G)

Figure 1: Example of an entry: BAAL ou BEL

## Developed system

Figure 2 shows the overall picture of the developed system. We develop a web application with Google's Firebase, which will solve one of the challenges mentioned earlier, the difficulty of collaboration. To create the data in the left part of the figure, we used the text available on Wikisource. Some of the authority information is mechanically extracted in advance, using the rules for extracting authority information obtained in Volume 1. As of the end of May, 4374 entries for Volume 2 are registered. The developed system, shown in the center of the figure, provides functions such as input form for authority information, user management, and review function. The text is displayed on the left side of the input screen, and the input form is displayed on the right side. The user enters the necessary information in the form on the right side, copying the text displayed on the left side as necessary. This input assistance function contributes to reducing input errors, which was one of the issues mentioned above. It also provides functions to highlight the relevant part of the text based on the authority information entered and assists the normalization of notational errors in text based on the achievement of Volume 1. The registered data will be used for the visualization of the number of registered entries. In addition, it can be exported in CSV format, which increases the independence of data from the system, and leads to the long-term preservation of data.



**Figure 2: Overview of the developed system**

## Conclusion

This research describes the developed system to support the extraction of bibliographical data from the Encyclopédie. Three months after the system launched, the number of registered users is 16, and they entered 484 bibliographical data out of 4374 entries, which falls short of our expectations. Therefore, we should encourage the participation of a broader range of people, not just members of the Society, by improving the system

functionality as follows: (1) to provide functions for users who are not good at French, with which they can read the 18th-century French text while referring to online dictionaries and machine translation; and (2) to adopt gamification and visualization techniques that automatically display and update the network of entries and authors based on the inputted bibliographical data. Such techniques would contribute to upkeep the participants' motivation and share the results beyond entering bibliographical data.

## Bibliography

- Académie des Sciences (2017). Édition Numérique Collaborative et CRitique de l'Encyclopédie. <http://enccre.academie-sciences.fr/encyclopedie/>.
- Department of Romance Languages and Literatures of Univ. of Chicago ARTFL Encyclopédie. <https://encyclopedie.uchicago.edu/>.
- Earhart, A. E. et al. (2020). Citational politics: Quantifying the influence of gender on citation in Digital Scholarship in the Humanities. *Digital Scholarship in the Humanities*, (fqaa011). doi:10.1093/lc/fqaa011.
- Fiormonte, D. (2015). Towards Monocultural (Digital) Humanities? *Infolet*. <http://infolet.it/2015/07/12/monocultural-humanities/>.
- Hashimoto, Y. and Kano, Y. (2019). Honkoku2: Towards a Large-scale Transcription of Pre-modern Japanese Manuscripts. *Proceedings of the 9th Conference of the Japanese Association for Digital Humanities*. pp. 97–100.
- Kawashima, K. (2005). *Émilie Du Châtelet to Marie Lavoisier: 18 Seiki France No Gender to Kagaku (Émilie Du Châtelet and Marie Lavoisier: Gender and Sciences in the 18th-Century France)*. Tokyo: Univ. of Tokyo Press.
- Murai, H. and Tokosumi, A. (2007). Network Analysis of the Four Gospels and the Catechism of the Catholic Church. *The Journal of Advanced Computational Intelligence and Intelligent Informatics*, 11(7): 772–79. <http://www.bible.literarystructure.info/2007SCISISIS.pdf>.
- Nagasaki, K. et al. (2017). A Collaborative Approach between Art History and Literature via IIF. 12th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2017, Conference Abstracts. <https://dh2017.adho.org/abstracts/185/185.pdf>.
- Partner of Pelagios Network. *Recogito*. <https://recogito.pelagios.org/>.
- Romanello, M. (2016). Exploring Citation Networks to Study Intertextuality in Classics. *Digital Humanities Quarterly*, 010(2). <http://www.digitalhumanities.org/dhq/vol/10/2/000255/000255.html>.
- Schwab, Richard N. with the collaboration of Walter E. Rex. (1971-72). *Inventory of Diderot's Encyclopédie*, 6 vols, SVEC, 80, 83, 85, 91-93. Oxford: The Voltaire Foundation.
- Yoshiga, N. (2021). Conversion of Historical Ogihan Business Records into Linked Open Data via Human-Machine Cooperation. In *The National Museum of Japanese History* (ed), *Japanese and Asian Historical Resources in the Digital Age*. Golden, CO: Fulcrum, pp. 121–45. <https://hdl.handle.net/2027/fulcrum.5d86p217p>.

# Platformed reflections on the Pandemic: Covid-19 and Electronic Literature

Anna Nacher<sup>1</sup>, Søren Bro Pold<sup>2</sup>, Scott Rettberg<sup>3</sup>

## Introduction

The paper reflects on the experience of collaboratively curating the online exhibition of digital art and electronic literature, [1] “COVID e-lit”, which opened in May 2021 as part of Electronic Literature Organization 2021 Conference & Festival. The project, supported by a small DARIAH-eu grant, incorporated a series of 13 interviews with artists featured at the exhibition carried out in February and March of 2021, which were consecutively used as material for the documentary to have its premiere in June 2021.

## Approach

The sheer fact that born-digital creative output is prone to ephemerality and technical obsolescence informs a specific auto-reflexivity of digital media, manifesting among other ways as a meaningful, sustained, and consistent effort at archiving potentially fluid and unstable content. This requires a significant amount of infrastructure and several communities of practice that have emerged as a way to archive rich, technologically, and aesthetically robust e-literary content. Through our project, located at the crossroads of documentation, research, and archiving, we tap into a vigorous archiving activity of e-literature scholars, practitioners, and audiences. One of the major outlets is ELMCIP Knowledge Base, where we are setting up a research collection [2] “Pandemic E-literature” in order to establish a basis for comparing different approaches to documenting and processing the pandemic through digital art.

---

<sup>1</sup> Jagiellonian University

<sup>2</sup> Aarhus University

<sup>3</sup> University of Bergen

## Pandemic E-Lit

### Research Collection

#### Collection curated by:

Scott Rettberg  
Søren Bro Pold  
Anna Nacher

#### Description:

This is a research collection used to collect works and critical writing that are reflective of the COVID-19 Pandemic. This research collection is part of the "Electronic Literature (e-lit) and Covid-19 Research" by Anna Nacher, Søren Pold, and Scott Rettberg, funded by Dariah-EU.

#### Creative works:

Title ▲	Author	Year
Content Moderator Sim	Mark Sample	2020
Coronary (Coronário)	Giselle Beiguelman	2020
Coronation: a webcomic	Marino Family, Mark C. Marino	2020
Curt Curtal Sonnet Corona	Amaranth Borsuk	2020
Ear for the Surge	Claire Fitch	2020
Exposed	Sharon Daniel, Erik Loyer	2020

**Figure 1: Research Collection at ELMCIP Knowledge Base**

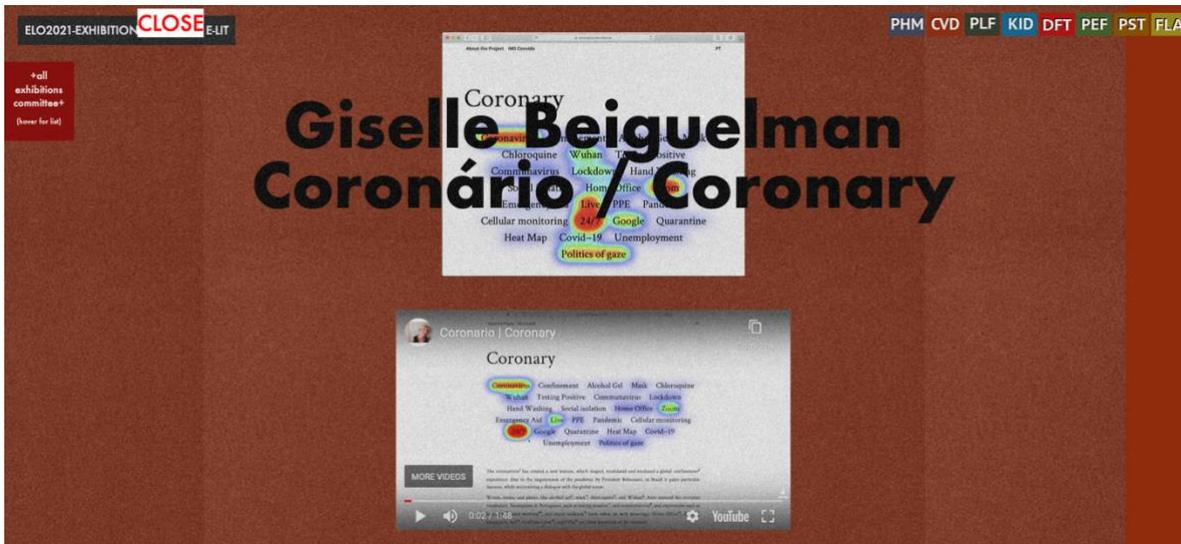
The impulse to archive is all the more meaningful in the age of the pandemic when in the course of barely a couple of months in 2020 we witnessed almost every aspect of our everyday life transferred online. This quickly resulted in both an explosion of new forms of creativity and the acute syndrome of new forms of platform overload such as the much-touted “Zoom fatigue”. The former were often mediated through popular platforms and brought new digital genres in addition to already recognized but still under-researched ones. During the COVID-19 crisis, digital networked platforms took center stage and corporations like Facebook, Google, Apple, and Zoom became our main interfaces to public space and cultural life. The affordances of popular digital services not only shaped our responses and reactions to the crisis but the algorithms they are based on also reveal their [3] “residual power” (Parisi, 2013: 13), increasingly [4] posing significant challenges to human perception and cognition (Clough, 2018). One of the results from such a shift is related to what Naomi Klein has named “Screen New Deal”, where commercial technology platforms are introduced to a much larger extent and integrated into our private spaces and homes. Digital artists are developing new forms of critical reflection on this new arrangement of our relationship to technology, e.g. about how platforms invade privacy, intimacy, individual spaces and simultaneously impact collaboration, communities, cultural and political life. Such reflection often takes the form of digital

artwork, viewable online, available for download, or functioning as a browser plug-in, that can serve as critical digital media.

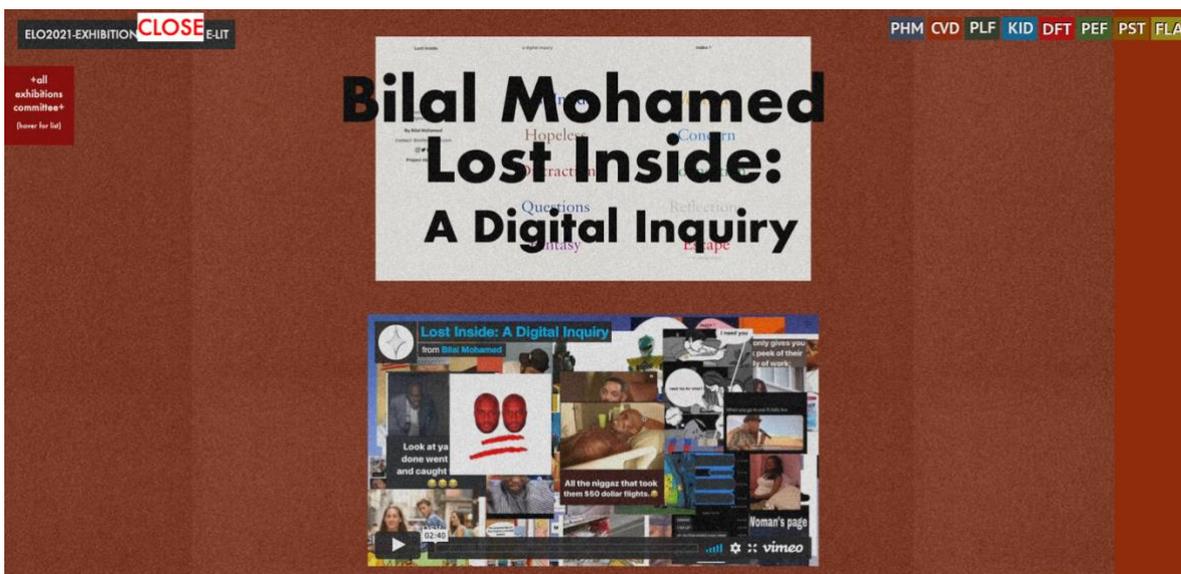
When we started discussing the Covid-19 pandemic and electronic literature, we reflected on the fact that, while there are many public memorials related to wars, there are very few related to epidemics and diseases. Apart from the horrible scenes we have seen from hospitals around the world, the everyday experience of the pandemic has for many of us mostly been visible through its lack of normality, as reflected for example in closed down, deserted cities. The Covid E-lit exhibition portrays all this through art and electronic literature as, we hope, an already historical document of life under the pandemic. We believe the exhibition demonstrates that the pandemic, besides all its horrors and cancellations, has also been a genuine moment for art and electronic literature. In the paper, we analyze how such a strategy of digital artwork functions both as a public memorial - developed both as fluid and unstable digital artwork and as plethora of documenting and archiving activities emerging around it - and as a performative critique of digital culture through self-reflective networked media amplified by the pandemic. Both these facets have emerged across diverse sites of creative interrogation and production, exemplifying what Luciana Parisi calls [3] “aesthetic function of algorithms” resulting in “outbreak of randomness within logic” (Parisi 2013: xv). The latter can be also ascribed to the instances of creative use of platforms and datastreams contrasted with their normative, prescriptive and habituated popular deployment. The process has accelerated with the massive COVID-19-induced shift online, thus signalling the possible prospect of a significant change of paradigm.

Based on the insights gained through various aspects of this practice-based and ethnographic approach, our analysis explores the cultural moment of the COVID-19 pandemic through the lens of digital culture as reflected in electronic literature and digital art produced in sites that differ geographically, socially, and culturally. The method allowed us a glimpse into various aspects of the early stage of the pandemic in diverse locations. From a net art visual essay on cultural experience of pandemic life in Brazil, where President Bolsonaro obfuscated its very existence (Giselle Beiguelman’s *Coronario / Coronary*), to a Muslim-Oromo artist’s attempt at journalling elusive yet visceral sense of isolation resulting in deepened relationship with himself and the world at large (Bilal Mohammed’s *Lost Inside: A Digital Inquiry*), to documentation of overwhelming scope and scale of COVID-19-related humanitarian crisis happening in spaces rarely explored or even brought to attention of wider public: the American infamous criminal punishment system, overcrowded and unsanitary even well before the pandemic (Sharon Daniel’s and Erik Loyer’s *Exposed*). Some of the artworks explicitly emphasize the troubling relationship between affective response to the overwhelming news on virus transforming

all the social sphere within barely a few days and the profit- and data-mining oriented commercial platforms, such as Twitter or Facebook (Ben Grosser’s Doomsscrolling, and to some extent Mark Sample’s Content Moderator Sim).



**Figure 2: Giselle Beiguelman, Coronario / Coronary (Covid E-lit Online Exhibition)**



**Figure 3: Bilal Mohammed, Lost Inside: A Digital Inquiry (Covid E-lit Online Exhibition)**

### Conclusion

In sum, the exhibition demonstrates how digital art, embedded and operating within platforms, can provide a unique take both on platform culture and the experience of living within platforms during a time of global crisis.

**Reference**

Rettberg, S., Pold S. B., Nacher A. (2021). *Pandemic E-lit. Research Collection*. ELMCIP Knowledge Base, <https://elmcip.net/research-collection/pandemic-e-lit> , Accessed 15.08.2021.

Rettberg, S., Pold S. B., Nacher A. (2021). *Covid E-lit online exhibition*. <https://www.eliterature.org/elo2021/covid/> , Accessed 15.08.2021.

Parisi, L. (2013) *Contagious Architecture. Computation, Aesthetics, and Space*. Cambridge, Mass: MIT Press.

Clough, P. T. (2018). *The User Unconscious. On Affect, Media, and Pleasure*. London: University of Minnesota Press.

## Digital Humanities and the way forward for ethnographic research: What we learned from Covid-19?

Deepika Kashyap<sup>1</sup>

### Abstract

The ongoing Covid-19 pandemic has put many academic disciplines and research institutions at risk and has created an uncertain future as many of the ongoing research, or the proposed research are disrupted or stopped indefinitely. Many scholars, academicians, and researchers grapple with the thoughts and techniques to overcome this unprecedented event and resume their work safely. Now, the safety of the researchers and the subjects are ethical and moral questions for many institutions and disciplines. Although all academic disciplines are affected by this pandemic, the degree of intensity varies from discipline to discipline. There are disciplines like anthropology, sociology, ethnology, and folklore whose laboratories are located outside the four walls of their research institute, which are adversely affected by the pandemic. Researchers from the mentioned disciplines closely deal with the everyday lives of people where intimacy and closeness are used as part of the research methodologies, and field notes are considered the holy text. The researchers try to become an active part of their research subject to provide a “thick description” of the field; in other words, a detailed description of social practices and behaviors of the people is necessary. So, what happens when we remove the “field” and the “people” from the disciplines like anthropology, sociology, ethnology, and folklore. How do they carry forward their research without going to the field? As we can see, many scholars from the field of anthropology, sociology, ethnology and folklore are ready to think beyond the physical ethnography ever before. Now, the researchers are more conscious of the vulnerabilities of ethnographic research (Corey M Abramson, 2020), and the current pandemic has forced them to think about an alternative method. As many anthropology and ethnology departments suggest their scholars and researchers to blend their research and adopt the online research methodologies or digital ethnography (DE) to resume their research during the pandemic, they really don't address the methodological issues lies in the digital ethnography. How do we make notes from digital ethnography? How do we study and analyze the data collected through digital ethnography? Do we have any specific tool/tools to analyze the notes collected from digital ethnography?

As a scholar of ethnology and folklore, I will elucidate my experience with digital ethnography during the Covid-19 and how it changed my perspectives towards

---

<sup>1</sup> Department of Communication, University of Hyderabad, India  
17snpc06@uohyd.ac.in

physical ethnography. While doing the digital ethnography, the foremost question that comes to mind is how to approach the digital data and analyze them because manually analyzing the digital data is a lengthy process; also, it is not the same as our field notes. So, how do we negotiate with digital ethnography? Do we have any digital tool/tools to study digital ethnography? Of course, we can smoothly study digital ethnography using digital humanities (DH), and I will illustrate it through my own engagement with DH and how it helped me study folklore on the online platform.

In the internet and social media age, it is impossible to ignore the viral videos, memes, stories, and fake news on our newsfeed. This digital transformation of material culture has opened new directions in folklore research by transcending its physical boundary. Now one can equally rely on the data available in the virtual world without really bothering about the field notes. The only matter of concern is the tools and techniques involved in studying online data/information. Also, we cannot ignore the concept of “thick description” as we are dealing with people and their everyday lives and behaviors. Since the internet is a vast source of information, the folkloric material available on the internet is astounding. So, instead of focusing on big data or a large amount of data, it is important to have “thick data” (Wang 2013) and descriptive analysis of the data. We can have the thick data by implementing DH in our DE as I used Voyant Tools for text mining in my research on folkloric material on the Facebook page. Voyant Tools is one of the important tools to extract the meaning of the text and find a pattern on the texts posted on the Facebook page. Since hundreds of people posting countless posts on the Facebook page and there are thousands of comments on each post, it is difficult for the researcher to analyze data manually. So text mining method is time-saving and efficient to analyze the big data; at the same time, it keeps the concept of “thick description” intact. Besides the data collection and analysis process, DH is more transparent since there is no human feeling or perspectives involved in it. We can better understand this debate of transparency in ethnographic research through a comparative study between physical ethnography and digital ethnography.

This paper will shed some light on both physical and digital ethnography since my digital journey is an extension of my physical encounter with people, folklore, and their culture. Before the Covid-19 pandemic, my research was based on my field notes and active participation in the field, so it is important to understand the influence of pandemic and natural calamities on ethnographic research and how DH can build a way forward for the ethnographic research.

# Virtual Communities and Post-Pandemic Possibilities: Animal Crossing New Digital Humanities

Quinn Dombrowski<sup>1</sup>, Elizabeth Grumbach<sup>2</sup>, Merve Tekgürler<sup>3</sup>

## Introduction

The absence of sustained opportunities to network, collaborate, and communicate in the physical world due to the onset of the coronavirus pandemic has necessitated virtual conferences and events in our profession, but platforms like Zoom and Twitter don't fully capture the embodied experience of community. Inspired by the affordances of video game platforms to simulate real worlds, the Animal Crossing New Digital Humanities (ACNDigHum) lecture series<sup>4</sup> is an experiment in designing a more embodied and inclusive platform for virtual networking, research communication, and community building during the pandemic. Reflecting on both the lecture series and the special half-day livestream during the 2020 U.S. Presidential Election, the authors will present on the successes and challenges of these events, how the virtual Animal Crossing world can inspire real-world change in the ways we conduct professional events, and recommendations for approaching community-building in a post-pandemic world that will demand hybridity and accessibility.

## Approach

Animal Crossing New Horizons (ACNH, released March 2020) is a life simulation game that allows players to create an avatar that reflects their real-world physical attributes and gender expression. While many of the personal customization options were present in previous versions of Animal Crossing (AC), ACNH is the first game in the series to allow the player to change their skin and hair color. In addition to building a house that can be customized with elaborate flooring, wallpaper, and items, users are allowed to terraform an entire island and create themed spaces using representations of real-world objects. The game has few big-picture goals and objectives, and so ongoing engagement with the game is spurred by personal goals and possibilities enabled by the passage of time (and tied to the real-world calendar). Compared to the previous AC games, where character customization was limited to a few ready-made options, ACNH allows for almost

---

<sup>1</sup> Stanford University

<sup>2</sup> Arizona State University

<sup>3</sup> Stanford University

<sup>4</sup> *Animal Crossing: New Digital Humanities*, <https://digitalhumanities.stanford.edu/acndh>, Accessed on Aug. 8, 2021.

unlimited self-expression, allowing players to feel “present” in the virtual world that they have a direct hand in developing.

The ACNDigHum lecture series drew inspiration from DH academics sharing their ACNH experiences on Twitter as lockdowns began across the globe. When the ADHO DH 2020 conference shifted to online, one of the ACNDigHum creators worked with Shawn Moore to host a series of lightning talks. An in-game social event took place, as well, recreating familiar conference spaces (e.g. conference rooms with sponsor paraphernalia). The success of this pilot event inspired the ACNDigHum creators to undertake an ongoing talk series beginning in October 2020.

The talk series is publicly viewable using the Twitch platform, which is widely used to broadcast live-streamed content. Anyone can visit the ACNDigHum Twitch URL<sup>5</sup> from a web browser and ask questions using Twitch’s chat functionality. There is a higher technical barrier to entry for presenters and “in-person” attendees, who must have Nintendo Switch, a copy of ACNH, and a Nintendo Online subscription. They use their own console to fly their avatar to the presentation location using a one-time-use “Dodo Code”. Voice audio for “in-person” participants is captured using a separate app, Discord, which is commonly used for voice and text chat, particularly among gaming communities. One of the ACNDigHum hosts uses their laptop to capture video from ACNH, combine it with audio from Discord, and broadcast it to Twitch with minimal lag.

Over the past year, DH groups and organizations have quickly adapted events to online formats. Without the constraints of physically co-existing in the same real-world space, scholars have faced an overabundance of events they could attend on Zoom. However, there has been a limited amount of experimentation with alternate platforms, such as CSDH-SCHN adopting Gather.Town, but even this relies on video camera and microphone audio, with their concomitant complications around privacy.

The ACNDigHum space is also reflective of the unavoidable porousness of personal and professional boundaries during the pandemic. The “DH hangout space” exists on an island with four human players, two of them younger than 8. Even in the absence of children, ACNH itself is not designed to be a passive, invisible space. Insects run across the ground, the weather can change hourly, and non-player characters may decide to stroll directly in front of a human player giving a presentation. While video conferencing environments seem like spaces independent from usual distractions of everyday life, where only the faces of the speakers are visible and the use of a virtual background is common to avoid outside distractions, AC is designed with these distractions in mind. The game

---

<sup>5</sup> “ACNDigHum” on Twitch, <https://www.twitch.tv/acndighum>, Accessed on Aug. 8, 2021.

interacts with the players, breaking with the fiction of digital spaces that assume personal life can be separated from the professional while working from home.

The use of a game platform like Animal Crossing has invited scholars from a variety of fields to take up more creative means of presenting their work than the simple slide deck. While scholars have argued that indie game development should be considered a form of digital humanities, ACNDigHum demonstrates the viability of using games to talk about a wide variety of scholarship, hosting talks from fields including English, East Asian literature, Ottoman Turkish history, book studies, Black Studies, and medieval studies (Ruberg, 2018; Coltrain and Ramsay, 2019). Only one presentation involved any use of slides; others have used in-game designs, items, and avatar clothing changes (sometimes into custom designed clothing items that depict visualizations) to illustrate points.

## Conclusion

Scholars in game studies have argued that video games can lead to “potentially significant long-term social change” and be harnessed to solve real-world problems (Bogost, 2007; McGonigal, 2011). In this paper, we argue that video games and other interactive, virtual platforms can and should inspire digital humanities to find new methods of community-building, especially as we move into a new phase of globally uneven pandemic recovery that will require us to adapt beyond the challenges of the past year. Using ACNH as a lens, and the ACNDigHum lecture series as a case study, this paper will present possible methods for creating more hybrid, accessible, and inclusive environments for scholarly knowledge communication in the digital humanities.

## Reference

- Bogost, Ian** (2007). *Persuasive Games: The Expressive Power of Video Games*. Cambridge, MA: MIT Press.
- Coltrain, James and Stephen Ramsay** (2019). “Can Video Games Be Humanities Scholarship?” In Gold, M. and **Klein, L.** (eds), *Debates in the Digital Humanities 2019*. U of Minnesota Press, Accessed on August 20, 2021. <https://dhdebates.gc.cuny.edu/read/untitled-f2acf72c-a469-49d8-be35-67f9ac1e3a60/section/10c2899a-d78c-40d2-b293-f828d3a1b3e9>.
- McGonigal, Jane** (2011). *Reality is Broken: Why Games Make Us Better and How They Can Change the World*. New York: Penguin Press.
- Ruberg, B.** (2018). “Queer Indie Game Making as an Alternative Digital Humanities.” *American Quarterly* 70(3): 417-438.

# Building Web Corpus of Old Nubian with Interlinear Glossing as Digital Cultural Heritage for Modern-Day Nubians<sup>1</sup>

So Miyagawa<sup>2</sup>, Vincent W.J. van Gerven Oei<sup>3</sup>

## 1. Old Nubian

Old Nubian is a Nubian language recorded in the kingdoms of Nobadia and Makuria in the Middle Nile Valley, modern-day southern Egypt and northern Sudan, between the 8th and 15th centuries CE<sup>4</sup>. Besides Meroitic, the Old Nubian language is the oldest written language from the Nilo-Saharan language phylum<sup>5</sup>. It was written using the Coptic-derived Nubian alphabet, and includes three characters from the Meroitic alphasyllabary. The Old Nubian corpus consists of both literary material, such as Bible translations and sermons, and burial texts, and documentary materials, such as contracts, land sales, as well as numerous wall inscriptions of various kinds.

## 2. Goal of the project

Despite recent advances in the description of Old Nubian language and the publication of numerous new texts, there is no central digital corpus of Old Nubian texts. A team consisting of So Miyagawa, a digital humanist and Coptologist, and Vincent van Gerven Oei, the current expert in the Old Nubian language, have started to compile a digital corpus of Old Nubian. The underlying goal is to make Old Nubian texts available via a linguistically and philologically tagged corpus of Old Nubian that can be accessed via a user-friendly, highly visual online digital platform. Data sustainability and interoperability will be realized using *de facto* digital humanities standards; for instance, TEI XML for the linguistic and philological tag mark-up of Old Nubian texts. A visualization will also be created that is helpful for both linguistic experts and Nubian heritage holders. Specifically, the team has opted to show interlinear glosses following Leipzig Glossing Rules<sup>6</sup> (LGR;

---

<sup>1</sup> This work was supported by JSPS KAKENHI Grant Numbers JP20K21975, JP21K00537.

<sup>2</sup> Center for Studies of Cultural Heritage and Inter Humanities (CESCHI), Kyoto University

<sup>3</sup> Punctum Books and Community-led Open Publication Infrastructures for Monographs (COPIIM), Coventry University.

<sup>4</sup> For more information on the genealogy and history of Old Nubian and Nubian languages, see Chapter 1 of van Gerven Oei 2021 [1].

<sup>5</sup> Though it is doubtful whether Nilo-Saharan forms a coherent genetic unity, the genealogical relations among Eastern Sudanic languages within Nilo-Saharan language phylum, which include Old Nubian and Meroitic, are well-established. See Rilly 2010 [2].

<sup>6</sup> The *de facto* standard for interlinear glossing in linguistics papers, created by the Department of Linguistics at the Max Planck Institute for Evolutionary Anthropology, Leipzig [3].

modified to suit Old Nubian) under each word and morpheme on the web corpus for professional linguists. In addition, the corpus will in the future also incorporate Arabic and plain English grammatical annotations for Nubians, heritage holders of Old Nubian, especially those without a linguistic educational background.

### 3. Corpus-building methodology

The Old Nubian corpus data were primarily provided by Vincent van Gerven Oei. The data were originally composed in XeLaTeX, and the interlinear glossing were realized in a gb4e.sty format that consists of four lines: the first line is the Old Nubian text in the Old Nubian alphabet, the second line is the Romanized version of the same Old Nubian text with the morphemes parsed by hyphens, the third line provides interlinear glossing, and the fourth line is the English translation of the Old Nubian texts.

The entire Old Nubian corpus is not as big as that of Coptic and far smaller than that of Ancient Greek; however, it contains large amounts of epigraphic materials. The most sizable text is Pseudo-Chrysostomus's *In venerabilem crucem sermo*, a ~3,200-word Old Nubian translation of a Greek homily. The second most sizable text is the *Miracle of Saint Mina*, a Nubian miracle story of ~950 words. As the project is currently in the commencement phase, we are creating a pipeline to automatically convert the interlinear glossed corpus of Old Nubian from LaTeX or CSV into an adaptation of the TEI XML format developed for the Coptic SCRIPTORIUM<sup>7</sup> and subsequently web pages with the interlinear glossing remaining under the text.

### 4. Pilot project

First, we chose the *Stauros Text* as the prototype. At ~800 words, the *Stauros Text* is a relatively long text in comparison to the other Old Nubian texts. Vincent van Gerven Oei developed the interlinear glossed text file in XeLaTeX which he is currently rendering into a CSV spreadsheet. To convert this into a simple interlinearly glossed text with an English translation in this LaTeX file, So Miyagawa created an XSLT program that converted the LaTeX file into TEI XML format, which was shared on GitHub to support data interoperability. An example of the TEI XML file that resulted from the conversion between LaTeX and gb4e.sty is below (Fig. 1).

---

<sup>7</sup> The first multi-layered corpus project of Coptic texts [4]. For the project, see Schroeder and Zeldes 2016 [5].

```

<ab xml:id="SC4">
  <s type="orig">Γαειᾶ οὐκ ὀκιδᾶρρε·</s>
  <s type="parse">Γαει-ᾶ οὐ-κ ὀκ-ιδ-αρ-ρ-ε</s>
  <s type="roman">ηaei-a ou-k ok-ij-ar-r-e</s>
  <s type="gloss">who-QUOT 2PL-ACC call-PLACT-INTEN-PRS-1SG.PRED</s>
  <s type="trans" xml:lang="en">‘What shall I call you?’</s>
  <note>Notes The following affirmative forms in -μα are all dependent on the verb
    ὀκιδᾶρρε.</note>
</ab>

```

**Figure 1: Tiers tagged by <s> </s> are the original text, parsed text, Romanized parsed text, interlinear glossing and English translation in TEI XML.**

He then created a further XSLT program to transform the TEI XML into an HTML file that was combined with a JavaScript file to enable visualization of our interlinear glosses in the form of LGR on any Internet browsers (Fig. 2). This pipeline produced the first online Old Nubian corpus.

A large part of the documentary material is yet to be digitized and is unavailable in the XeLaTeX format described above. We will work on a pipeline that will allow efficient data entry for the remaining materials.

Γαειᾶ οὐκ ὀκιδᾶρρε·		
Γαει-ᾶ	οὐ-κ	ὀκ-ιδ-αρ-ρ-ε
ηaei-a	ou-k	ok-ij-ar-r-e
who- <u>QUOT</u>	<u>2PL-ACC</u>	call- <u>PLACT-INTEN-PRS-1SG.PRED</u>
‘What shall I call you?’ <sup>[2]</sup>		

**Figure 2: Interlinear corpus rendition of Fig. 1 as a web page transformed by XSLT and leipzig.js [6].**

## 5. Future perspectives for digital cultural heritage

Using the pipeline described above, we will produce more Old Nubian literary texts with LGR-styled interlinear glossing. To facilitate an online search, we will create a search function of each lemmata and glosses on the homepage using XQuery and also provide photos of Old Nubian manuscripts in IIF if their affiliations permit us. Moreover, as a side-product from the interlinear glosses, we intend to create the first digitized lexicon data of Old Nubian for further NLP development for Old Nubian.

Our work is primarily aimed at Old Nubian philology experts or linguistics, or historians in Medieval Nubia. However, we are also planning to embed plain English and Arabic translations, easy-to-read explanations for each gloss, a basic grammar of Old Nubian, and lecture videos on our corpus homepage. Through doing so, we hope to contribute to the cultural preservation and heritage education of Nubians and spread knowledge of the Nubian culture to a wider global audience.

## Reference

- [1]. Vincent W.J. van Gerven Oei, *A Reference Grammar of Old Nubian*, Leuven: Peeters, 2021.
- [2]. Claude Rilly, *Le méroïtique et sa famille linguistique*, Leuven: Peeters, 2010.
- [3]. Max Planck Institute for Evolutionary Anthropology - Department of Linguistics, Leipzig Glossing Rules: Conventions for Interlinear Morpheme-by-Morpheme Glosses, <https://www.eva.mpg.de/lingua/resources/glossing-rules.php>, accessed on 2021-05-05.
- [4]. Caroline T. Schroeder, Amir Zeldes, et al., Coptic SCRIPTORIUM: Digital Research in Coptic Language and Literature, <https://copticcriptorium.org/>, accessed on 2021-05-05.
- [5]. Caroline T. Schroeder and Amir Zeldes, Raiders of the Lost Corpus, *Digital Humanities Quarterly* 10 (2), 2016, <http://www.digitalhumanities.org/dhq/vol/10/2/000247/000247.html>, accessed on 2021-06-21.
- [6]. Benjamin Chauvette, Leipzig.js: Interlinear Glossing for the Browser, <https://bdchauvette.net/leipzig.js/>, accessed on 2021-06-21.

# Development of data-driven historical information research infrastructure at the Historiographical Institute in the University of Tokyo

Satoru Nakamura<sup>1</sup>, Taizo Yamada<sup>1</sup>

## Introduction

The Historiographical Institute in the University of Tokyo has developed a digital archive called “SHIPS” with about 40 different databases (total of 5.6 million data) and about 20 million images of historical materials from ancient times to the Meiji Restoration, including a database of historical catalogs, the full-text data of historical materials, images of historical materials, and tools such as kuzushiji dictionary for historical materials and historical information on Japanese history. By inheriting and developing the past efforts of the Historiographical Institute, we start to develop an information environment infrastructure that can sustain the accumulation of 150 years of data over the next 100 years. Specifically, we will build (1) a data repository that enables long-term data accumulation and access, (2) a data-driven search system that maximizes the use of the accumulated data, finally (3) building a foundation for historical informatics research to strengthen our international dissemination capabilities. In this research, we will describe our efforts to build a data-driven search system.

## Approach

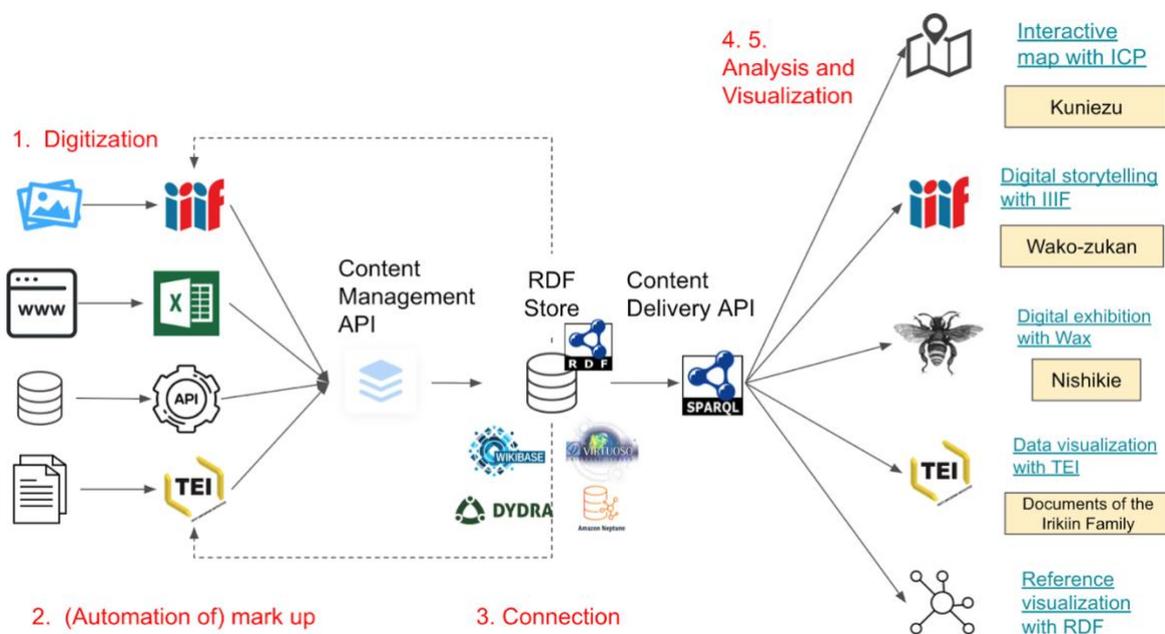
The construction of a data-driven infrastructure for historical information research is a theme that has been actively pursued both in Japan and abroad. In Japan, ROIS-DS Center for Open Data in the Humanities (CODH) is conducting historical big data research [2], and In Europe, the Time Machine Project [1] is conducted. The "Time Machine Project" aims to build a large-scale simulator for European history and to convert the vast collections of cultural institutions into a digital information system. In addition to the results of research in the humanities and history, the following technological elements are required to realize this project [3]; (1) Digitization, (2) Automation of markup, (3) Connection, (4) Simulation engines, and (5) Experience.

While referring to the technological elements of the "Time Machine Project," this study defines the following five technological elements as necessary for the construction of a "data-driven historical information research infrastructure," positioning step 4 and step 5 as processes for utilizing accumulated data; (1) Digitization, (2) Automation of markup, (3) Connection, (4) Analysis, and (5) Visualization.

---

<sup>1</sup> The University of Tokyo

Figure 1 shows the overview of the system we are developing based on this methodology. Regarding the first point, the development of low-cost and high-quality digitization technology and the creation of solutions for the long-term preservation of digital data are required. The Historiographical Institute has already been actively digitizing its archives and will continue its efforts to digitize and preserve them for the long term in cooperation with other projects, such as the Program for Constructing Data Infrastructure for the Humanities and Social Sciences [4]. As for step 2, by combining the results of research in the humanities and history with AI/machine learning, we develop the methodology and system to generate structured text by such as extracting the names of people and places with named entity recognition. For step 3, we structure data using international standards such as IIF and TEI, and furthermore, Linked Data/RDF, which is a technology for relating structured data, can be used to improve their interoperability. Step 4 is the process of creating new knowledge by utilizing the accumulated data with various technologies, such as text mining technology by Yamada et al [5]. Step 5 is the visualization of the results obtained in 4, which may include the use of maps and network diagrams.

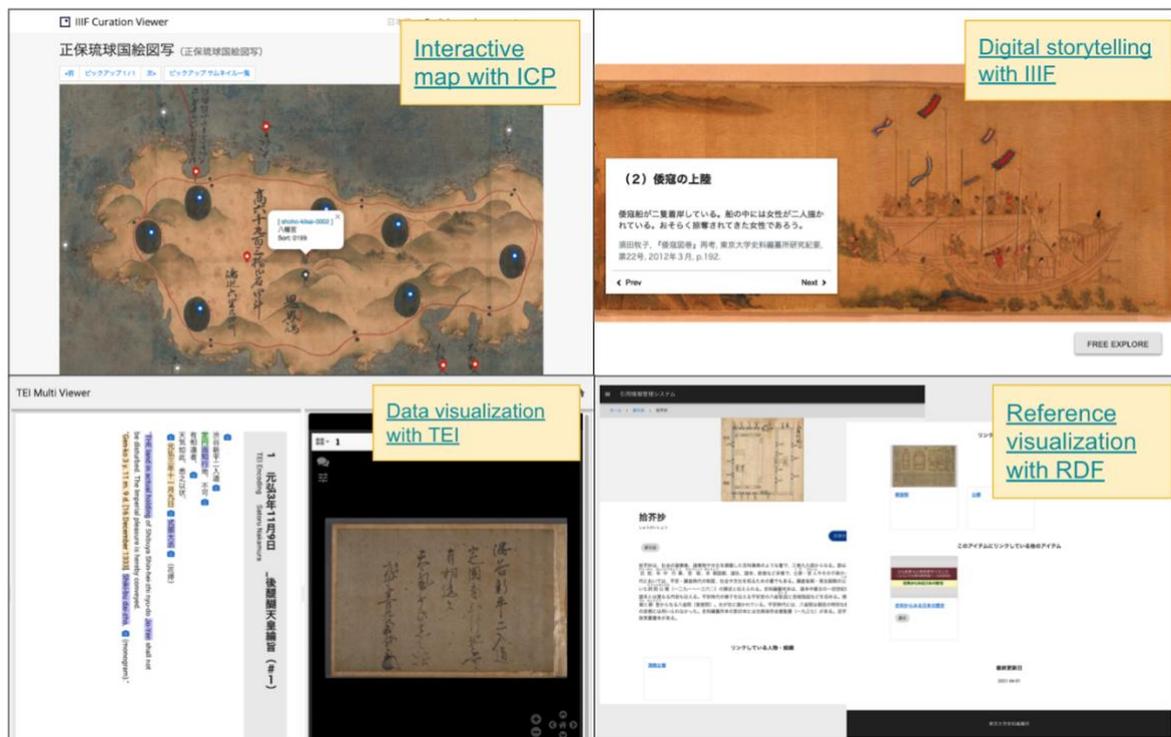


**Figure 1: Overview of approach for the construction of data-driven historical information research infrastructure.**

### Application examples

Figure 2 shows an example of the application we are currently developing. In the upper left of the figure, we are working on creating an interactive map of Kuniezu, using ICP (IIF Curation Platform) and the Edo map [6] as a reference. The upper right figure shows an

application that uses IIF Presentation API 3.0 and Canvas Panel [7] to realize storytelling about the contents of a picture scroll, such as Wako-zukan. In the lower-left of the figure, text data is structured using TEI and enables to display text and images in parallel, and unique expressions in the text are associated with the knowledge base. The lower right of the figure shows an application that visualizes the relationship between historical documents used in past exhibitions and the names of people and places that appeared in their captions. Structuring this kind of information leads to extract and accumulate the knowledge needed for future data association.



**Figure 2: Example of applications we are currently developing.**

## Conclusion

In this research, we report on our activities to build a data-driven infrastructure for historical information research. At present, we are in the process of preparing data for the development of data-driven functions (structuring data using IIF, TEI, and RDF). In the future, we will collaborate with experts inside and outside of the Historiographical Institute, and work on data association and knowledge base construction to support the efficiency of historical research and various outputs.

## Reference

- [1]. Time Machine Europe., <https://www.timemachine.eu/>, Accessed on 2021-06-05.
- [2]. Asanobu KITAMOTO and Mika ICHINO and Chikahiko SUZUKI and Tarin CLANUWAT, Historical Big Data: Reconstructing the Past through the Integrated Analysis of Historical Data, Eighth Conference of Japanese Association for Digital Humanities (JADH2018), pp.67-69, 2018.
- [3]. Harry Verwayen, The European Context, IIF Conference 2019, <https://youtu.be/8KWM36wY-QM>, Accessed on 2021-06-05.

- [4]. Japan Society for the Promotion of Science, Program for Constructing Data Infrastructure for the Humanities and Social Sciences, <https://www.jsps.go.jp/english/e-di/index.html>, Accessed on 2021-06-05.
- [5]. Taizo Yamada, Satoshi Inoue, Collecting Name of Historical Person from Historical Material Related to Japan, Digital Humanities 2017, 2017.
- [6]. Asanobu KITAMOTO and Shoko TERAOKA and Misato HORII and Hiroshi HORII and Chikahiko SUZUKI, Integrating Historical Maps and Documents through Geocoding - Historical Big Data for the Japanese City of Edo, Digital Humanities 2020, 2020.
- [7]. Digirati - Cultural Heritage | Canvas Panel, <https://cultural-heritage.digirati.com/building-blocks/canvas-panel/>, Accessed on 2021-06-05.

# Compilation of Semantic Data Archive: A New Method of Learning “Local Culture”

Kwangwoo Kim<sup>1</sup>, Soohyeon Kim<sup>2</sup>

## 1. Introduction

In October 2020, Mokpo-si<sup>3</sup> of South Korea and the Center for Digital Humanities at the Academy of Korean Studies<sup>4</sup> launched a database compilation project called **Construction of Mokpo Modern History Archive**<sup>5</sup> to accumulate data and information on Mokpo City's modern history with various forms of digital data such as text, image, geographic information, and 3D models. The large-scale research project, which is conducted by Kim Hyeon, professor at the Academy of Korean Studies, includes a variety of detailed research projects such as a survey and 3D modeling of modern architectural heritage, and collection and archiving of related historical records.

This paper examines the structure and content of the Mokpo Modern History Archive implemented as one of the project's detailed tasks, and examines how implementing such a semantic database can contribute to local cultural research and education, and what challenges need to be improved.

KEY WORDS: Mokpo City of South Korea, Modern Architectural Heritage, Semantic Database, Local Culture and History

## 2. Resources for Archive Data Compilation

In 1894, a treaty was signed between Korea and Japanese to open Mokpo City as a trade port. Under this treaty, Mokpo was opened on October 1, 1897,<sup>6</sup> The fact that the U.S., France, Russia, Germany, Britain and Japan were involved in the provisions of the treaty signed at the time of the opening of Mokpo shows that Mokpo was a city where people

---

<sup>1</sup> 1st Author, Center for Digital Humanities, The Academy of Korean Studies, exfinder@naver.com

<sup>2</sup> Corresponding Author, Center for Digital Humanities, The Academy of Korean Studies, clayart141@naver.com

<sup>3</sup> Mokpo-si is a city with a population of 220,000 located on the southern coast of South Jeolla Province, Korea.

<sup>4</sup> Center for Digital Humanities at the Academy of Korean Studies <http://dh.aks.ac.kr>

<sup>5</sup> Construction of Mokpo Modern History Archive <http://dh.aks.ac.kr/~mokpo/wiki/index.php>

<sup>6</sup> 徳間一芽, 2010, Construction of a town by Japanese immigrants and their urban lives in Mokpo during the open-port period in Korea, Chonnam National University, 13p

from various countries could live. However, since the annexation of Japan and Korea in 1910, most foreigners in Mokpo have been Japanese, and many of the modern architectural heritages related to those Japanese people have remained in Mokpo.

The research team collected data on the modern history of Mokpo City from 85 local or central institutions such as Mokpo City Hall, Mokpo City Cultural Center, Regional Newspaper and Broadcasting companies in Mokpo, National Library, Korea Newspaper Archive, and National Institute of Korean History.

### 3. Archive Design and Data Compilation

The Archive was designed to make all the information elements be linked to each other, to implement a semantic data archive that can visually show the relationship.

The ontology for implementing data archive used the ekc Data Model <sup>7</sup> (Data Model for the Encyclopedic Archives of Korean Culture) developed by the Center for Digital Humanities at the AKS.

6 humanities researchers (majored in history, literature, and philosophy) who can understand and analyze the content of the historical records participated in the creation of semantic data in accordance with the ontology design. Kim Kwang-woo and Kim Soohyeon, the authors of this paper, are humanities majors and digital humanities researchers who acquired data processing technology, so they were able to play a central role in building the semantic database.

The design of the data archive and the numbers of the data nodes collected and organized under this design are as follows:

#### 3-1. Data Classes and Amount of Data

The schema of the Archive was designed with 9 classes including Actor, Event, Place, Architecture, Object, Record, Concept, Digital Asset, and Web Resource<sup>8</sup>. The amount of data currently implemented for each class (until 10 August 2021) is shown in the table below.

Class	Sub Class	description	Nodes (2021. 8. 10.)
Actor	<a href="#">person</a>	people who were involved in historical events of Mokpo City.	201

<sup>7</sup> EKC: [http://dh.aks.ac.kr/Encyves/wiki/index.php/EKC\\_Data\\_Model-Draft\\_1.1](http://dh.aks.ac.kr/Encyves/wiki/index.php/EKC_Data_Model-Draft_1.1)

<sup>8</sup> Ontology Class: <http://dh.aks.ac.kr/~mokpo/wiki/index.php/Ontology:Class>

	<u>group</u>	organizations, institutions, administrative agencies	187
Event		historical events, commemorative events.	2,129
Place		places in the geographic and administrative system related to historic buildings or historical events	192
Architecture		historic buildings chosen for 3D modeling	76
Object		artworks, artifacts, tools, monuments, museum objects	30
Record	<u>literature</u>	books, magazine, newspaper, documents	86
	<u>text</u>	article, text in historical records	3,210
	<u>multimedia content</u>	old photographs, sound record	4,040
	architectural drawing	drawing, floorplan of buildings	480
Concept		concepts, terms to explain historical facts	100
Digital Asset	<u>3D Model</u>	3D modeling data of historic buildings	73
	<u>360° surround VR</u>	aerial and landscape views of the historic places of Mokpo city	192
	<u>still image</u>	photographs of the present details of historical buildings and places	1,100
	<u>video clips</u>	Interview records, visual documentary	175
Web Resource		explanatory text from the Encyclopedia of Korean Culture, Korean Cultural Heritage Portal, etc.	400
total			12,671

### 3.2 Relation Design and RDF Data Implementation

Approximately 12,000 nodes compiled in the data archive are being described as RDF(Resource Description Framework) data that explicitly expresses the relationship between data nodes using the object property vocabulary defined by the ekc Data Model.

As the work to create RDF data is currently underway, about 6,000 links were created by August 2021, and a total of 1,800 RDF statements will be created by October 2021, including 12,000 semantic relations, 6,000 relations between semantic data nodes and multimedia assets or web resources. Object property vocabulary terms for RDF data creation are as follows.

NameSpace	Relation	Inverse Relation
dcterms:	A creator B	B isCreatorOf A

ekc:	[s] A writer B	B isWriterOf A
ekc:	[s] A calligrapher B	B isCalligrapherOf A
ekc:	[s] A inscriber B	B isIncriberOf A
ekc:	A translator B	B isTranslatorOf A
ekc:	A annotator B	B isAnnotatorOf A
ekc:	A founder B	B isFounderOf A
ekc:	A constructor B	B isConstructorOf A
ekc:	A reconstructor B	B isReconstructorOf A
ekc:	A renovator B	B isRenovatorOf A
dcterms:	A contributor B	B isContributorOf A
dcterms:	A publisher B	B isPublisherOf A
dcterms:	A rightsHolder B	B isRightsHolderOf A
edm:	A isDerivativeOf B	
edm:	A isSuccessorOf B	
ekc:	A hasOldName B	B isOldNameOf A
ekc:	A isNamesakeOf B	B isEponymOf A
ekc:	A administrates B	B isAdministratedBy A
ekc:	A participatesIn B	B hasParticipant A
ekc:	A documents B	B isDocumentedIn A
ekc:	A goesWith B	
ekc:	A isUsedIn B	
edm:	A isNextInSequence B	B isPreviousInSequence A

ekc:	A performed B	B isPerformedBy A
ekc:	A isPerformedAt B	
ekc:	A hasExhibitionAt B	
edm:	A happenedAt B	
ekc:	A depicts B	B isDepictedIn A
ekc:	A mentions B	B isMentionedIn A
dcterms:	A references B	B isReferencedBy A
ekc:	A isSteleOf B	B hasStele A
ekc:	A isStupaOf B	B hasStupa A
ekc:	A isEnshrinedIn B	B enshrines A
edm:	A currentLocation B	B isCurrentLocationOf A
edm:	A formerLocation B	B isFormerlocationOf A
dcterms:	A provenance B	B isProvenanceOf A
ekc:	A hasWife B (=isHusbandOf)	B isWifeOf A (=hasHusband)
ekc:	A hasConcubine B (=isHusbandOf)	B isConcubineOf A (=hasHusband)
ekc:	A hasSon B (=isFatherOf)	B isSonOf A (=hasFather)
ekc:	A hasSon B (=isMotherOf)	B isSonOf A (=hasMother)
ekc:	A hasDaughter B (=isFatherOf)	B isDaughterOf A (=hasFather)
ekc:	A hasDaughter B (=isMotherOf)	B isDaughterOf A (=hasMother)
ekc:	A hasAdoptedHeir B	B isAdoptedHeirOf A
ekc:	A hasBrother B	
ekc:	A hasSister B	

ekc:	A hasSonInLaw B (=isFatherInLawOf)	B isSonInLawOf A (=hasFatherInLaw)
ekc:	A hasSonInLaw B (=isMotherInLawOf)	B isSonInLawOf A (=hasMotherInLaw)
ekc:	A hasDaughterInLaw B (=isFatherInLawOf)	B isDaughterInLawOf A (=hasFatherInLaw)
ekc:	A hasDaughterInLaw B (=isMotherInLawOf)	B isDaughterInLawOf A (=hasMotherInLaw)
ekc:	A hasDescendant B	B isDescendantOf A
ekc:	A isLineageKinOf B	
ekc:	A isAffinalKinOf B	
ekc:	A hasDisciple B (=isMasterOf)	B isDiscipleOf A (=hasMaster)
ekc:	A hasOwner B	B isOwnerOf A
ekc:	A hasSubject B	B isSubjectOf A
ekc:	A servesAs B	
ekc:	A wasOrdainedBy B	B wasPreceptorOf A
foaf:	A knows B	
ekc:	A isFellowOf B	
dcterms:	A hasPart B	B isPartOf A
foaf:	A member B	B isMemberOf A
owl:	A sameAs B	
ekc:	A isNear B	
ekc:	A wears B	B isWornBy A
dcterms:	A type B	
edm:	A isRelatedTo B	B isRelatedTo A
edm:	A isShownAt B	





**Figure 2: Examples of 3D Model Data: Honam Bank Mokpo Branch and Japanese Consulate Mokpo<sup>11</sup>.**

## 5. Conclusion

Semantic databases create multi-directional links between fragmented information elements and allow users to explore broader, deeper knowledge. In addition to providing digitized research resources, this archive of the modern history of Mokpo is expected to serve as a local cultural education content that allows students and citizens of Mokpo City to understand the history and culture of their hometown in more depth. Citizens may be able to develop new cultural activities based on their understanding of culture in the history of their hometown obtained through this archive.

This semantic data archive is a database with various advantages and values of utilization compared to conventional bibliographic data archives. However, from the perspective of researchers who participated in this database compilation project, there are a number of challenges that need to be improved to increase the potential values of this archive.

The most deficient aspect of the current archive is that many of the data nodes provide only basic information and the related details do not exist. For example, a Japanese resident who first built or owned a building that remained a modern cultural heritage of Mokpo City is known only for his name and occupation, but no more information is available. Such information may be more likely to be found in the remaining records in Japan than in Mokpo or Korea.

In order to make the knowledge information in this archive richer and more valuable, I think that Japanese researchers interested in the history of modern cities in East Asia should take their data and participate in this research project together. The archive also expects to provide a digital research environment to Korean and Japanese scholars where they can discover common interests and conduct collaborative research activities.

<sup>11</sup> 3D Model Data: [http://dh.aks.ac.kr/~mokpo/wiki/index.php/3D\\_Model\\_Home](http://dh.aks.ac.kr/~mokpo/wiki/index.php/3D_Model_Home)

## 6. Reference

- [1]. Center for Digital Humanities at the Academy of Korean Studies <http://dh.aks.ac.kr>
- [2]. 徳間一芽, 2010, Construction of a town by Japanese immigrants and their urban lives in Mokpo during the open-port period in Korea, Chonnam National University, 13p.
- [3]. EKC: [http://dh.aks.ac.kr/Encyves/wiki/index.php/EKC\\_Data\\_Model-Draft](http://dh.aks.ac.kr/Encyves/wiki/index.php/EKC_Data_Model-Draft)
- [4]. Ontology Class: <http://dh.aks.ac.kr/~mokpo/wiki/index.php/Ontology:Class>
  
- [5]. Honam Bank Network graph. [http://dh.aks.ac.kr/cgi-bin/encyves/Story02.py?db=s\\_okehkim&project=mokpo&key=호남은행](http://dh.aks.ac.kr/cgi-bin/encyves/Story02.py?db=s_okehkim&project=mokpo&key=호남은행)
- [6]. Explain about Honam Bank Network graph: <https://youtu.be/alKoqdbszm4>
- [7]. 3D Model Data: [http://dh.aks.ac.kr/~mokpo/wiki/index.php/3D\\_Model\\_Home](http://dh.aks.ac.kr/~mokpo/wiki/index.php/3D_Model_Home)

# Towards a Structured Description of the Contents of the Taisho Tripitaka

Yoichiro Watanabe<sup>1</sup>, Kiyonori Nagasaki<sup>2</sup>, Hyunjin Park<sup>3</sup>, Yifán Wáng<sup>4</sup>, Tomohiro Murase<sup>5</sup>, Masayoshi Watanabe<sup>6</sup>, Norimichi Yajima<sup>7</sup>, Yoshihiro Sato<sup>3</sup>, Yūi Sakuma<sup>3</sup>, Xinxing Yu<sup>3</sup>, Masahiro Shimoda<sup>1</sup>, Ikki Ohmukai<sup>1</sup>

## 1. Introduction

We, the SAT Daizōkyō Text Database Committee, maintain a database of the e-texts of the Buddhist scriptures contained in the Taisho Tripitaka, which contains 2920 Buddhist scriptures and over 100 million kanji characters, and have made them available to the public in searchable form. However, these text data are based on tagged e-texts produced in 1994, and are not yet fully structured. For this reason, we are in the process of converting e-texts of the Taisho Tripitaka into the TEI format. In this presentation, we would like to introduce a part of our markup strategy in accordance with the TEI guideline.

## 2. Problems with the existing text

First of all, I will describe the problems of the existing e-texts. Conventional tagged e-texts, which we maintain, do not have tags to indicate detailed content divisions such as "chapters", but only tags to indicate volume divisions, which are easier to identify from the outside. The current SAT system, which is based on traditional tagged e-texts, reflects only the volume classification but does not go into the content classification. The current SAT system is also based on traditional tagged e-texts, which reflects only volume classification but does not go into content classification. Therefore, the current SAT system makes it difficult for the general public to search for, for example, only the chapter 方便品

(Chapter of Skillful Means) in the Lotus Sutra. Therefore, there is a need for e-texts that can express semantic hierarchy for researchers. In addition, although CBETA (中華電子佛典協會) has published its own extended TEI format that allows mark-up of content classifications, and e-texts based on this format, it was necessary to find a TEI format without its own extensions for further versatility.

## 3. Strategy on content markup

---

<sup>1</sup> University of Tokyo

<sup>2</sup> International Institute for Digital Humanities

<sup>3</sup> Graduate school of University of Tokyo

<sup>4</sup> Graduate school of University of Tokyo, International Institute for Digital Humanities

<sup>5</sup> Council for the Promotion of Tripitaka Research

<sup>6</sup> Jodo Shu Research Institute

<sup>7</sup> Waseda University

The Taisho Tripitaka is divided into three sections: the Indo-Chinese section, the Chinese section, and the Japanese section. The Japanese section contains many texts that differ slightly in style from the two sections, but our research group is attempting to create a hierarchical description of the e-texts in order to formulate a unified markup strategy for the entire collection. In particular, we would like to focus on how to describe the hierarchical nature of the text, such as the distinction between the broad sense of the text, which is not the part written by the editor of the Taisho Tripitaka, and the narrow sense of the text, which is the part excluding the preface, foreword, colophon, etc.

First of all, the broad text of the entire Taisho Tripitaka is indicated by `<div1 type="taisho_body">`. What is not included in this is the header given by the editor of Taisho Tripitaka. Within `<div1 type="taisho_body">`, elements corresponding to the main text in the narrow sense are marked with `<div2 type="body">`. On the other hand, a preface is defined as `<div2 type="preface">`, a foreword as `<div2 type="foreword">`, and a colophon as `<div2 type="colophon">`.

`<div2 type="body">` is used as a tag to describe a semantic division corresponding to a "chapter", etc. If `<div2>` meant a chapter, and there was a subdivision corresponding to a subchapter, then `<div3 type="body">` is used for it. In turn, `<div4>`, `<div5>` etc. can be used to indicate the hierarchical structure.

The `@type` given to these tags is intended to identify whether the element is a body or not. The further meaning of the unit is described by the `@subtype`. For example, `<div2 type="body" subtype="品">` ("品" means chapter), in order to make it easier to understand the meaning of this unit for people. In addition, `<div2 type="body">` does not always represent a "chapter". For example, in the larger text Mahāratnakūṭa Sūtra (see Figure1), there is a superordinate category of 會 ("union") which unites the 品 ("chapter") . For this reason, `<div2 type="body" subtype="會">` `<div3 type="body" subtype="品">` is used in the Mahāratnakūṭa Sūtra.



Figure 1: The beginning of Union Six of the Mahāratnakūṭa Sūtra.

Finally, we would like to explain how the above strategy relates to the "volume" classification used in e-texts. As mentioned earlier, "volume" is not primarily related to semantic categories. In other words, a chapter often crosses several volumes. In our study group, we use the empty tag <milestone unit="fascicle\_beginning"/> to indicate the point at which a "volume" begins. As this is an empty tag, it has no text to be enclosed within it, and is used in the same way as the <pb/> tag for the start of a page, and the <lb/> tag for the start of a line. This is because the volume separator is only a physical division, and is considered to be an extension of the line and page separators. See Figure 2.

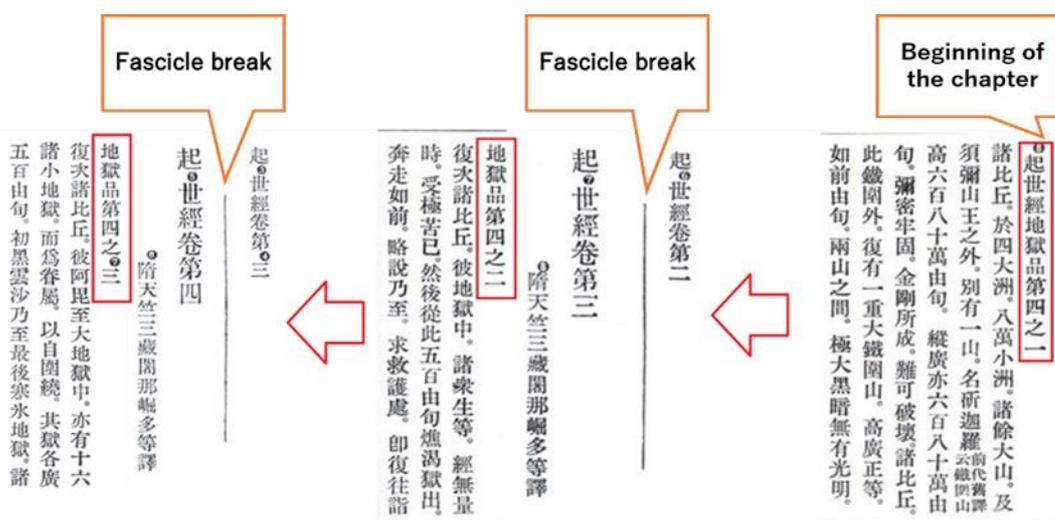


Figure 2: 『起世經』 (Qishi-jing) 「地獄品」 (chapter of hells) crosses three fascicles.

#### 4. Conclusion

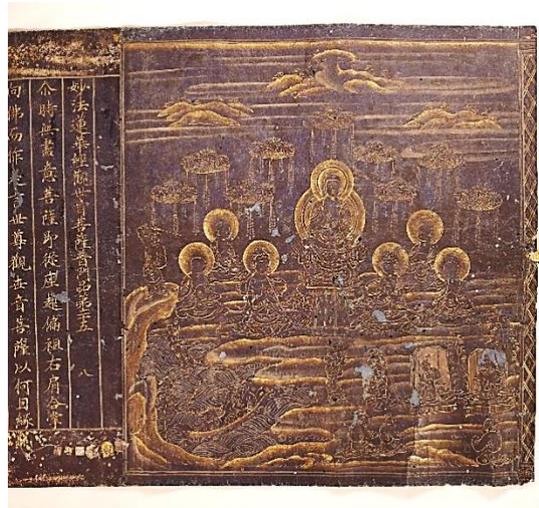
The creation of e-texts complete with these markup policies will be a meaningful contribution to the further internationalization and generalization of the TEI, not only in the narrow sense of Buddhist studies, but also as an example of the markup of classical Chinese materials. In the future, we aim not only to publish the TEI-formatted Taisho Tripitaka texts, but also to publish our mark-up strategy, so that the results of our research will be more universal and far-reaching. There are a number of issues that need to be addressed regarding the Japanese section, which we are currently working to resolve, but will report on soon.

# Classification of face images in the frontispiece paintings of Sutra copies in gold ink on indigo paper by deep convolutional neural networks

Toshiaki Aida<sup>1</sup>, Tomomi Kobayashi<sup>2</sup>, Aiko Aida<sup>3</sup>

The purpose of our research is to analyze the transition of the characteristics of the painting styles of Japanese manuscripts of Buddhist Sutra written in gold ink on indigo paper from the Heian period to the Muromachi period. In the Heian period, worship of the lotus Sutra flourished among court noblemen, and decoration of the sutra was very popular. Writing sutras with gold ink on indigo paper is one of the popular ways to decorate these sutras. Today many versions of this type remain extant. For example, the number of works designated as Japanese important cultural properties is around fifty sets [1]. A large part of them is the Lotus Sutra which made up of eight or ten scrolls, another large part is the Issai Kyo (the Tripitaka) which made up of about five thousand. Although many examples are preserved, they hardly have inscriptions on them, so their production backgrounds are still unclear. For this purpose, we utilize an excellent feature extraction ability of deep convolutional neural networks.

First, we have cut out about 200 face images of persons such as Buddha or Bodhisattva from among the images of the Frontispiece paintings in the old Japanese Buddhist Sutra copies that our research group has ever collected (Figure 1, 2). Then, we have put labels to them according to their production ages (Table 1). Although there are several works having inscriptions that indicates specific dates, many of them have neither inscription or record.

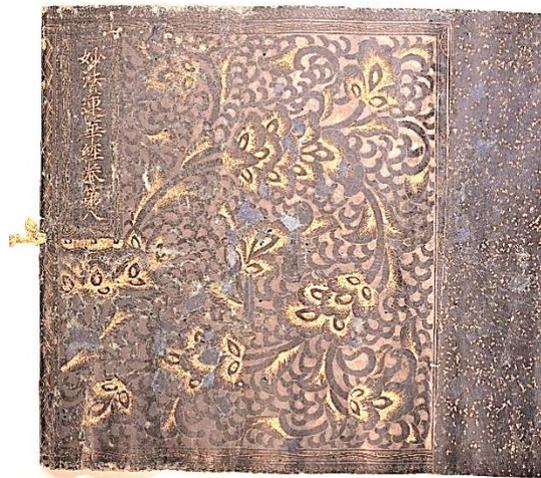


<sup>1</sup> Okayama University

<sup>2</sup> Chikushi Jogakuen University

<sup>3</sup> Japan Society for the Promotion of Science

**Figure 1: Frontispiece to fascicle eight of the Lotus Sutra, Fujii Eikan Bunko.**

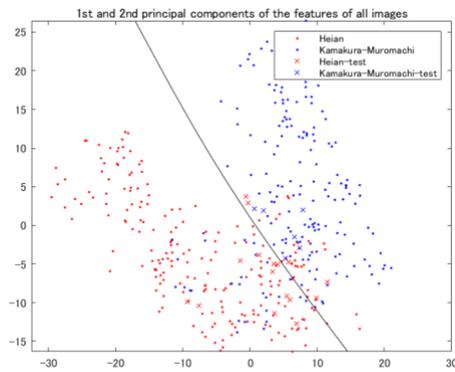


**Figure 2: Cover to fascicle eight of the Lotus Sutra, Fujii Eikan Bunko.**

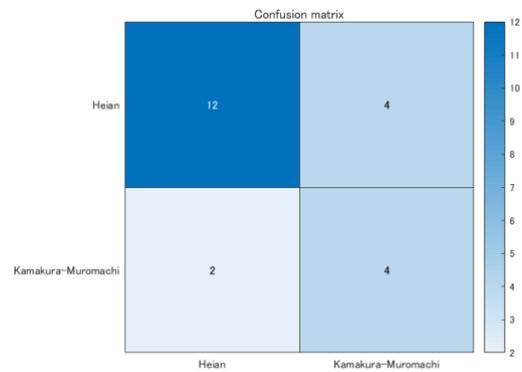
**Table 1: List of the works for dataset.**

No.	Name	Held by	Number of Scrolls	Hight (cm)	Width of Second Paper (cm)	Training/Test Data	Number of Faces	Period	Date of Surbey
1	Xiao zi jing (Chyuson-ji issai Kyo, in 1126)	National Museum of Japan History	1	25.4	44.6	Training Data	3	the Late Heian period	2018/04/19
2	Fascicle 87, Mahāprajñāpāramitāsāstra (Jingo-ji issai kyo, in 1155)	National Museum of Japan History	1	25.8	54.4	Training Data	3	the Late Heian period	2018/04/19
3	Vimalakirtinirdeśa (Jingo-ji issai kyo, in 1155)	National Museum of Japan History	2	25.9	55.7	Training Data	6	the Late Heian period	2018/04/20
4	Chyoju-o-kyo (Jingo-ji issai kyo, in 1155)	Iwase Bunko Library	1	25.8	52.6	Training Data	3	the Late Heian period	2011/02/20
5	Second fascicle of the Lotus Sutra	Fujii Eikan Bunko	1	25.6	50.1	Test Data E	3	the Late Heian period	2017/07/14
6	Sutra of Meditation on Samantabhadra Bodhisattva	Fujii Eikan Bunko	1	25.8	54.7	Test Data G	1	the Late Heian period	2017/07/28
7	Fascicle 169, Large Sutra on Perfect Wisdom	Fujii Eikan Bunko	1	25.5	55.8	Test Data E	3	the Late Heian period	2017/10/27
8	Lotus Sutra, Sutra of Innumerable Meaning and Sutra of Meditation on Samantabhadra Bodhisattva	Hyakusai-ji temple	10	26.3	55.5	Training Data	29	the Late Heian period	2007/01/23
9	Fascicle 4 and 8, Lotus Sutra	Kakurin-ji temple	2	25.5	49.1	Test Data F	6	the Late Heian period	2010/03/27
10	Fascicle 577, Large Sutra on Perfect Wisdom	Fujii Eikan Bunko	1	25.4	53.8	Test Data G	3	the end of the Heian period	2017/10/11
11	Fascicle 579, Large Sutra on Perfect Wisdom	Fujii Eikan Bunko	1	25.3	53.8	Test Data G	3	the end of the Heian period	2017/10/27
12	Fascicle 4, Lotus Sutra	National Museum of Japan History	1	24.7	45.1	Test Data E	2	the end of the Heian period	2018/05/07
13	Fascicle 1, 5, 6, 7 and 8, Lotus Sutra	Henmei-in temple	5	25.4	53.9	Training Data	15	the end of the Heian period	2009/06/14
14	Lotus Sutra, Sutra of Innumerable Meaning and Sutra of Meditation on Samantabhadra Bodhisattva	Kamigamo-jinja shrine	10	26.1	54.5	Training Data	29	the end of the Heian period	2019/03/29
15	Fascicle 355, Large Sutra on Perfect Wisdom	Kanagawa Prefectural Museum of History	1			Training Data	2	the end of the Heian period	2011/03/01
16	The Life Span of the Buddha	Kosetsu Museum	1	26.7	51.1	Test Data F	2	the end of the Heian period	2009/06/29
17	Lotus Sutra, Sutra of Innumerable Meaning and Sutra of Meditation on Samantabhadra Bodhisattva	Chyofuku-ji temple	10	26.8	47	Training Data	25	the Kamakura period	2010/03/16
18	Fragment of the eighth fascicle of the Lotus Sutra	Fujii Eikan Bunko	1	26	47.1	Test Data F	3	the Kamakura period	2017/11/02
19	Sutra of Innumerable Meaning	Fujii Eikan Bunko	1	26.1	44.9	Test Data E	3	the Kamakura period	2017/10/27
20	Fascicle 1, 2, 3, 5, 6, 7 and 8, Lotus Sutra (in 1304)	National Museum of Japan History	7	25.7	44.5	Training Data	19	the Kamakura period	2018/05/08
21	Fascicle 6, Lotus Sutra	Kanagawa Prefectural Museum of History	1	27	48	Test Data G	3	the Kamakura period	2011/03/01
22	Heart Sutra (in 1389)	Fujii Eikan Bunko	1	25.3	50.5	Training Data	3	the Muromachi period	2017/10/12
23	Lotus Sutra	Enzo-ji temple	8	26.3	56	Training Data	24	the Muromachi period	2010/03/18
24	Fascicle 2, 3 and 4, Lotus Sutra	Henmei-in temple	3	25.2	53.6	Training Data	9	the Muromachi period	2010/03/18
			Total:	72	scrolls	Total:		202	faces

Among the face images, focusing on the ones of the Late Heian, the End of Heian, the Kamakura and the Muromachi periods, we have made their feature extraction, applying a pre-trained deep convolutional neural network, ResNet101, pre-trained on the ImageNet dataset. Based on the features of the images, we have learned their classification, adopting support vector machine and principal component analysis as machine learning methods. As a result of 10-fold cross validation of the classification performance, the success rate more than 95% was found.



**Figure 3: 1st and 2nd principal components of the features of the images of the dataset F. Dots and crosses mean training and test data, respectively. A black curve is the boundary between the features of Heian(red) and Kamakura-Muromachi(blue) periods, which was inferred by support vector machine with its box constraint 0.1.**



**Figure 4: The confusion matrix for the dataset F. The diagonal and non-diagonal elements mean the numbers of successfully classified and misclassified test data, respectively.**

However, we have to note that the images of paintings by the same painter or atelier should not be included to both training and test data. Even if a painter depicts faces of different persons, excellent feature extraction ability of deep convolutional neural networks can detect that the faces were painted by the same painter. Therefore, if we included the images of paintings by the same painter to the both data, we incorrectly obtained a higher level of classification performance than their actual one, as if the deep neural networks had correctly learned the characteristics of the painting styles of the periods. Paying attention to the fact, we have prepared the training and test data, and tried a fairer performance evaluation of the classification of the images. As a result, we have found 82% of accurate rate of classification performance (Figure 3,4).

A similar study has been successfully attempted in order to detect a painter school for the paintings of the scenes from the Tale of Genji [2]. They made fine-tuning of a pre-trained deep convolutional neural network, and found which part of the paintings the fine-tuned network was focusing on to detect their painter school, using the Grad-CAM.

Furthermore, we have criticized and corrected, from the point of view of art history, the result of the inference of the production ages for test data, which we have made for performance evaluation of the classification by machine learning. The production ages of the Sutra in gold ink on indigo paper is evaluated comprehensively based on the

calligraphy of the paper, the design of the cover, and the materials and techniques used, as well as the paintings on the frontispieces. Therefore, the criteria do not necessarily focus on the painting style of the small faces. For example, in a previous study that included 44 plates of the Sutra Sutra in gold ink on indigo paper only four of them referred to facial expressions. In these works, the facial expressions of the late Heian period are briefly mentioned as "slightly full-cheeked face," "eyes and nose like hikime-kagihana; drawn-line eyes and a hook-shaped nose," and "gentle face" [3]. When the images of facial expressions misclassified by machine learning are reviewed by the human eye, the roundness of the cheeks is an important criterion, as has been pointed out (Figure 5, 6). On the other hand, for example, the length and slope of the line drawing indicating the eyes and eyebrows were also used as criteria in machine learning, even though the eyes and nose in the hooked-eye nose style looked the same at first glance.



**Figure 3: Misclassified Images for test data E (Fascicle four of the Lotus Sutra, National Museum of Japan History and Sutra of Meditation on Samantabhadra Bodhisattva, Fujii Eikan Bunko)**



**Figure 4: Misclassified Images for test data F (The Life Span of the Buddha, Kosetsu Museum, Fascicle 8 of the Lotus Sutra, Kakurin-ji temple and Fascicle 8 of the Lotus Sutra, Kakurin-ji temple, Fujii Eikan Bunko)**

As the usefulness of information technology has recently been shown in the field of art history [4], it is expected that expressions that have not been paid attention to in traditional study will be clarified as criteria for determining the period.

### **Acknowledgement**

This work was supported by a Joint Research Project of the Art Research Center, Ritsumeikan University and JSPS.KAKENHI (19J40241).

### **Reference**

- [1]. Agency for Cultural Affairs, "National Database of Designated Cultural Properties", <https://kunishitei.bunka.go.jp/bsys/index> (access data 2021/08/20)

- [2]. Takuya Kato, Mariko Inamoto and Akihiko Konagaya (2018), “Detection of a Painter School with Deep Learning Method: A Case Study of Scenes from the Tale of Genji,” The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2D1-05.
- [3]. Nara National Museum (1987). “Lotus Sutra – Calligraphy and Decoration”, Tokyo Bijyutsu, Tokyo, pp.440-471.
- [4]. Chikahiko Suzuki, Akira Takagishi, Alexis Mermert, Asanobu Kitamoto and Jun Homma (2020), “Analysis of difference between male and female facial expressions in Japanese picture scrolls using GM Method with IIF Curation Platform (a Poster Session),” The 10th Conference of Japanese Association for Digital Humanities (JADH2020, Proceedings), Osaka University (Online), Osaka, pp.90-95.

# The difference in transitional process between Western instrumental and vocal music

Daisuke Miki<sup>1</sup>, Akihiro Kawase<sup>2</sup>, Kenji Hatano<sup>2</sup>

## Introduction

The history of Western music is often discussed in the framework of a certain periodization. The terms Baroque, Classical, and Romantic are believed to represent the trait of music in a specific period. However, categories in a periodization cannot be explicitly contrasted because of the continuity and similarity among each period (Grout and Palisca, 1996).

Western art music is a complex consisting of instrumental music and vocal music. The development of vocal music has had a close relationship to the history of literature. One such example is the affinity of the poems by P. Verlaine (1844–1896) and the music by C. Debussy (1862–1918). Multiple studies were surveyed by Wright (2018), showing that the rhythm of a sung verse is closely related to the original text and that the richness of harmony reflects the emotional subtlety of the poem. Although music and literature share some of the historical divisions, such as the Baroque period and Romantic period, the period with an identical name in two areas of art did not necessarily emerge simultaneously nor in the same order. For this reason, the transitional process of vocal music such as opera, Lied, and *mélodie* should be considered different from how music without lyrics has developed.

Several studies have analyzed the stylistic evolution of music by focusing only on specific genres. Amino (1988) compared the structures of piano sonatas by W.A. Mozart (1756–1791) and L.v. Beethoven (1770–1827) investigating how many measures constitute each of the first and the second subject, exposition, development, and recapitulation. Daniele and Patel (2013) quantified the trend of historical change of German and Italian art song by computing the rhythmic feature called nPVI (normalized Pairwise Value Index). Shea (2017) focused on French and Italian opera and revealed the metric difference of opera across a timeline. However, those past studies are not consistent with the features with which they model stylistic evolution. Therefore, the achievements of these studies have remained unsynthesized into a comprehensive and objective analysis of music history.

## Aim of this study

---

<sup>1</sup> Graduate School of Culture and Information Science, Doshisha University

<sup>2</sup> Faculty of Culture and Information Science, Doshisha University

This study reveals the difference between Western instrumental and vocal music as to how each genre has altered its style across history. We analyzed instrumental and vocal music by 56 composers from J. Despres (1450–1521) to M. Ravel (1875–1937). According to Le Goff (2014), a periodization can only be applied to the civilization of a limited region. Therefore, we adopted the German, French, and Italian art music as the corpus to exclude an effect of what Le Goff (2014) called the globalism— a cultural delay due to the lag in exporting the culture of a specific region into a new world, which in the field of music, corresponds to America, Japan, or Slavic countries

## Methods

We acquired two sets of MIDI corpus, one of which consists of 281 instrumental music and the other of 853 vocal music. MIDI files were then converted into MusicXML, allowing us to extract musical features with which we carry out the quantitative analysis later on. We partitioned the samples of musical pieces with a k-means algorithm for each dataset using the variables that are interpretable from melodic, rhythmic, and harmonic point of view. The musical features adopted in the k-means clustering had been selected by Lasso (Least Absolute Shrinkage and Selection Operator) predicting the years of composition.

## Results and discussion

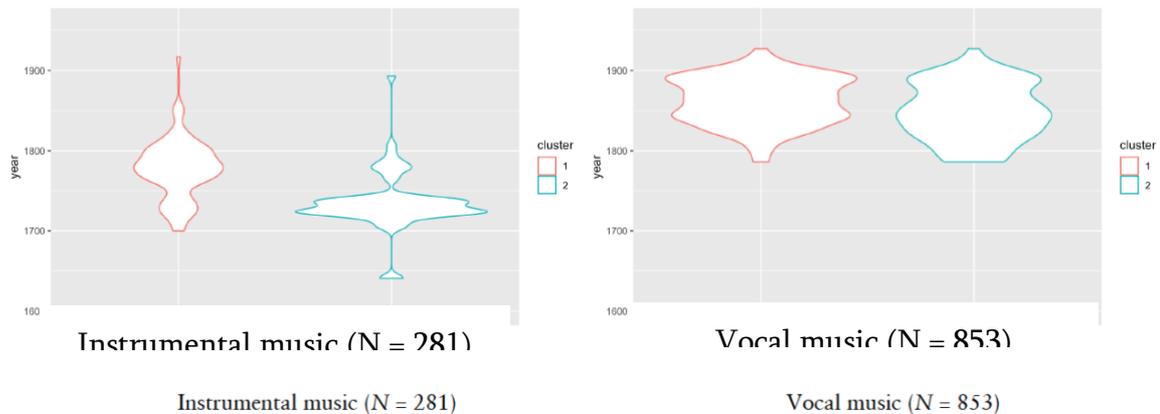
Lasso modeling zeroized the coefficients of the musical features that are not relevant to the time-shift, suggesting the musical features shown in Table 1 to be adopted for k-means clustering:

**Table 1: The musical features relevant to the time-shift (excerpt).**

	Instrumental music	Vocal music
Melodic features	$\pm 0 \rightarrow +12$ , $+1 \rightarrow +1$ , $+1 \rightarrow +6$ , $+14 \rightarrow \text{Rest}$ , etc.	$+2 \rightarrow \pm 0 \rightarrow \pm 0 \rightarrow +2$ , $+2 \rightarrow \pm 0 \rightarrow +3 \rightarrow -1$ , $2 \rightarrow \pm 0 \rightarrow \text{Rest}$ , etc.
Rhythmic features	Not relevant	$120 \rightarrow 120 \rightarrow 80 \rightarrow 80$ , $120 \rightarrow 360 \rightarrow 120$ $\rightarrow 240$ , $240 \rightarrow 120 \rightarrow \text{Rest} \rightarrow 120$ , etc.
Harmonic features	$\text{Ger}^7$ , $\text{I-IV-I-iv}$ , $\text{I-vii}^{\circ 7}$ , $\text{bII: I}$ , $\text{V}^7\text{-II-V}^7\text{-I}$ , $\text{i}^{\circ 7}\text{-V}$ , etc.	$\text{I-V-V}^7\text{/V}$ , $\text{I}^6\text{-I-vii}^{\circ 7}$ , $\text{II-V-I}$ , $\text{IV-V}^7\text{-I-V}$ , $\text{bIII: V}^7$ , etc.

The optimized number of cluster was computed in several ways (Elbow method, Gap Statistic, Silhouette Coefficient, and Canopy) and turned out to be  $k = 2$  for both vocal and instrumental music. We visualized the clustering result with violin plots with respect to the years of composition, and compared how the styles of the instrumental music and

the vocal music emerge, last, and eventually get replaced with their successive styles (Figure 1).



**Figure 1: Violin plots showing the stylistic shift in instrumental music and vocal music.**

This result indicates that the style of instrumental music changed dramatically around 1750 whereas the vocal music did not change its style throughout the last couple of centuries. However, we cannot ignore the bias in the range of the years of composition between each dataset. In order to enable more comprehensive and long-term stylistic evolution, we need to carry out further studies involving the vocal pieces in earlier centuries.

## Reference

- [1]. Amino, K. (1988). "Classicism and romanticism in art music: Analysis on the piano sonatas by Mozart and Beethoven" (in Japanese). *Gakushuin-Shigaku*, 26, pp.38–50.
- [2]. Daniele, J.R. and Patel, A.D. (2013) "An empirical study of historical patterns in musical rhythm: analysis of German & Italian classical music using the nPVI equation." *Music Perception*, 31(1), pp.10–18.
- [3]. Grout, D.J. and Palisca, C.V. (1996). *A History of Western Music* (5th ed.). New York: W.W. Norton & Co.
- [4]. Le Goff, J. (2014). *Faut-il vraiment découper l'histoire en tranches*. Paris: Editions du Seuil.
- [5]. Shea, N. (2107). "Meter in French and Italian opera." *Masters Theses*, 536, pp.1809–1859.
- [6]. Wright, A. (2018). "Research on the Compositions of Claude Debussy, Gabriel Fauré, and Reynaldo Hahn Set to the Poetry of Paul Verlaine." *MusRef*.

# **e-Sukhāvātī: An Innovative Digital Platform for Studying the *Smaller Sukhāvātīvyūha***

SIU Sai-yau<sup>1</sup>

## **Introduction**

Under the influence of the COVID-19 pandemic, university students have been studying at home through online platforms. In the past year, such a learning mode was adopted for my course on elementary Buddhist Sanskrit. To enable students to learn the ancient Indian language effectively, I have created a digital platform called “e-Sukhāvātī”. Focusing on the *Smaller Sukhāvātīvyūha*, a popular Mahāyāna scripture in Chinese Buddhism, the platform combines a sūtra reader with dictionaries, text analysis tools, and multimedia databases. It allows students to learn about essential concepts, cultural knowledge, and grammatical features of Mahāyāna Buddhist scriptures through a comparative study of the *Smaller Sukhāvātīvyūha* in Sanskrit and its Chinese translation by Kumārajīva (344-413), a renowned Buddhist translator in early medieval China. The web application aims to serve as an interactive learning tool helping students supplement what they learnt in online lectures and keeping them on track during remote teaching and learning.

## **Functions**

Entering the landing page of e-Sukhāvātī, users can access Kumārajīva's Chinese translation of the *Smaller Sukhāvātīvyūha* from the CBETA Online [1]. In addition, the system has been designed to be user-friendly, and all additional tools can be reached through a purpose-built search engine. The engine enables users to select a word or phrase in the text with a mouse, and a toolbar will appear on screen. They can then click on the tools to look for further explanations. Such a design minimises the time spent typing in search bars, speeds up the search for corresponding information, and makes the reading experience more enjoyable.

---

<sup>1</sup> Assistant Professor (School of Translation and Foreign Languages, The Hang Seng University of Hong Kong)





Figure 2: A digital corpus of the *Smaller Sukhāvāṭyūha* in Sanskrit and Chinese.

(2) Multilingual Buddhist dictionaries. In addition to NTI Buddhist Text Reader by Fo Guang Shan [3], e-Sukhāvati can be connected to DILA Glossaries for Buddhist Studies by the Library and Information Center at Dharma Drum institute of Liberal Arts [4]. It is a collection of authoritative Buddhist dictionaries in Pāli, Sanskrit, Chinese, Tibetan, English, and many other languages, providing free access to the public. Users can instantly gain access to the database by selecting a term from the sūtra with the built-in search engine. The platform will send the term to DILA Glossaries for further analysis and give a list of explanations from various Buddhist dictionaries.

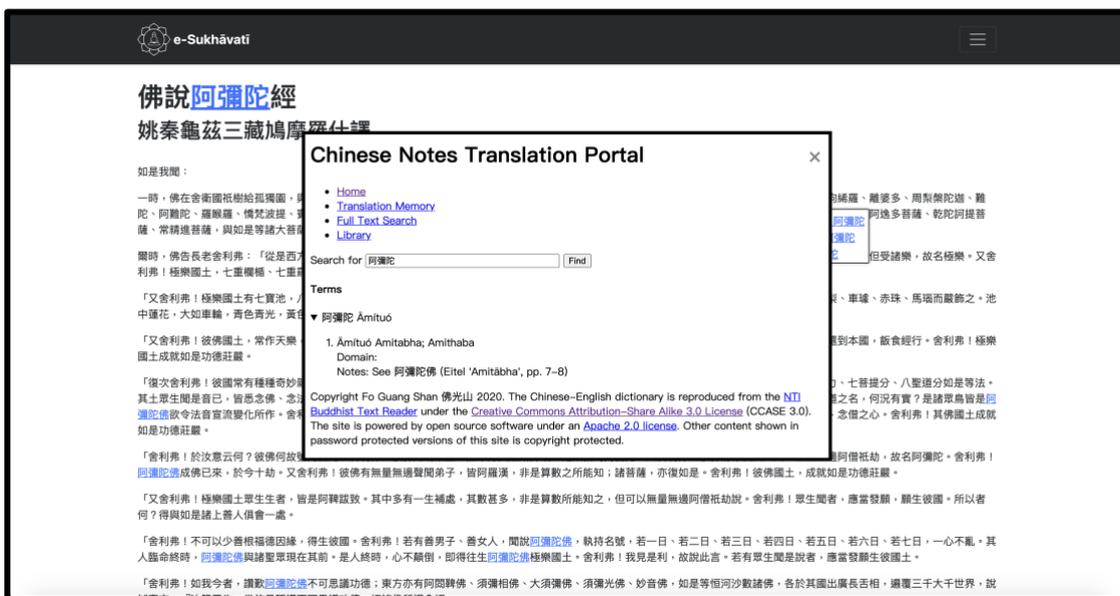


Figure 3: Explanations of a Buddhist term from NTI Buddhist Text Reader.

(3) A Sanskrit reader. The platform contains a recording of a recitation of the *Smaller Sukhāvātīvyūha* in Sanskrit. By clicking on the marker next to a section of the scripture, users can activate the audio file and listen to the Sanskrit recitation of the section. This is one of the unique functions equipped in e-Sukhāvātī.

(4) An AR-enabled multimedia database. To enhance the interactivity of e-Sukhāvātī and provide users with a comprehensive understanding of Amitābha's Western Pureland, by scanning the QR code on the platform with portable devices, users can view an animation of the Pureland and compare it with the descriptions in the scripture. In addition, the system is equipped with a variety of multimedia resources, such as 3D models of Amitābha and mural paintings of the Western Pureland in Dunhuang, by which users can familiarise themselves with the visualisation of the sacred realm in traditional art forms. Therefore, e-Sukhāvātī is intended for not only Sanskrit learners, but also students studying Buddhist art.



Figure 4: A 3D model of Amitābha on Sketchfab [5].

(5) An analytical tool for Chinese Buddhist terms. The toolkit contains a large collection of ancient Chinese texts from the Kanseki Repository [6]. When users enter a selected Buddhist term into the application, it will analyse the number of occurrences of the term in the ancient texts from the pre-Qin to Ming and Qing dynasties and display the statistics in an interactive line chart. The computer-aided analysis of textual data allows users to examine the spread of a Buddhist concept in ancient China by using theoretical frameworks from Reception Studies and Intellectual History. The analysis also helps explore the popularity of the concept and its underlying forces, facilitating the study of Buddhist scripture translation.

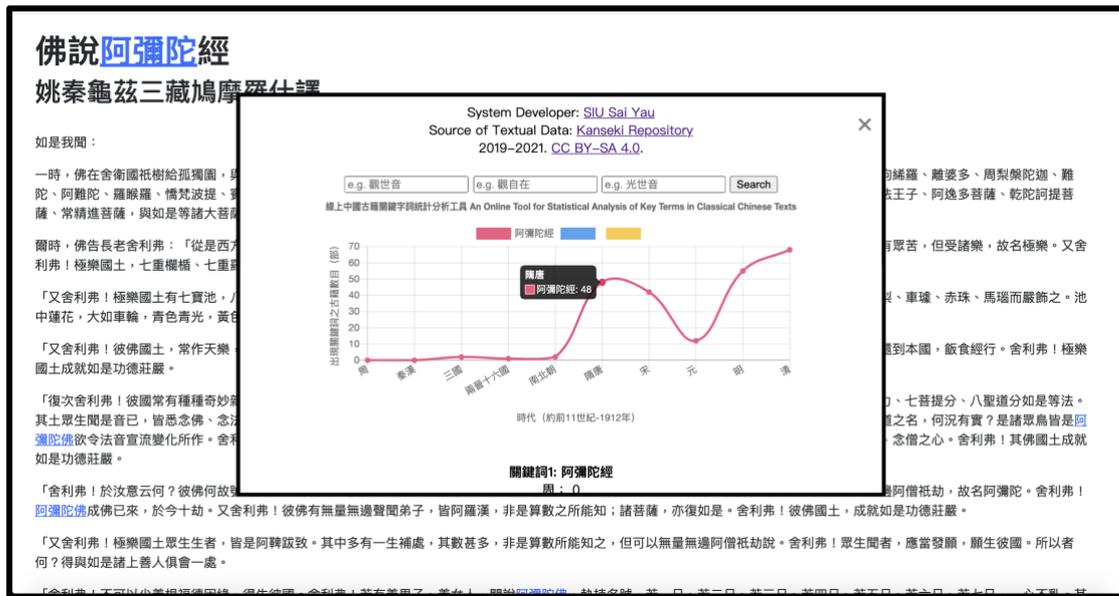


Figure 5: An analytical tool for Chinese Buddhist terms with data visualization.

### Future Development

There are three noteworthy points regarding the future development of e-Sukhāvātī: (1) Expanding the content. Currently, the platform contains only one Buddhist sūtra. In the future, I plan to develop the web-based application into “e-Tripiṭaka”, which will include more popular Mahāyāna Buddhist texts in Sanskrit and Classical Chinese, including *Prajñāpāramitāhṛdaya*, *Vajracchedikā Prajñāpāramitā*, *Vimalakīrtinirdeśa*, and *Saddharmapuṇḍarīka*. Subject to the availability of additional resources, digitised Buddhist scriptures in Pāli, Tangut, Mongolian, and Tibetan languages will also be incorporated into the new system; (2) Supporting the use of portable devices. At present, e-Sukhāvātī mainly works on desktop and laptop computers. The tool will be developed into a native app, enabling the public to use it on their tablets or smartphones and study the scriptures more easily; and (3) Incorporating AI-powered components into the platform. The system will be enhanced with a couple of additional toolkits developed by deep learning technology. The first one is an online multilingual translator, specifically designed to translate the sūtras into different languages. It will allow users to study the Buddhist scriptures in their native language. The second is a chatbot built with natural language processing and deep learning techniques, such as GPT-3. It will function as a virtual tutor, responding to users’ questions about the scriptures at all hours [7] [8].

### Reference

[1]. Digital Archive Section of the Library and Information Center, DILA. (2016). CBETA online [Computer software]. <https://cbetaonline.dila.edu.tw/zh/>

- [2]. Niedersächsische Staats- und Universitätsbibliothek Göttingen. (2008). The Göttingen register of electronic texts in Indian languages [Computer software].  
<http://gretil.sub.uni-goettingen.de/gretil.html>
- [3]. Fo Guang Shan. (2013). NTI reader [Computer software]. <https://ntireader.org/>
- [4]. Digital Archive Section of the Library and Information Center, DILA. (2016). CBETA online [Computer software]. <https://cbetaonline.dila.edu.tw/zh/>
- [5]. Marchal, G. (2016). Wooden sculpture of Amitābha [3D model on Sketchfab].  
<https://skfb.ly/NGMP>
- [6]. Kanseki Repository. (2014). 漢リポ Kanseki repository [Computer software].  
<https://www.kanripo.org/>
- [7]. Siu, S. C. (2013). Translation technology for the rendition of Buddhist texts: New directions in Buddhist translation. Saddharma Publishing House.
- [8]. Siu, S. Y. (2019). Buddhist scriptures: Contemporary translation and research methods. Saddharma Publishing House.

# **New Possibilities of Digital Publishing and Online Exhibition— A Case Study of the Website “Reflections on COVID-19”**

**Lin, Wen Jiun<sup>1</sup>**

## **Abstract**

This research, inspired by the Social Science Research Council's "Rapid-Response Grants on Covid-19 and the Social Sciences" program, represents Academia Sinica's contribution to addressing these issues.

Based on the digital publishing and exhibition result of the above program, the “Reflections on COVID-19” website represents the first stage of this enterprise, exemplifying the goals described above by publishing both academically and digitally. This website gathers the research of 19 Academia Sinica humanities and social science scholars, compiling their reflections on historical experiences and explorations of contemporary topics, as well as integrating a kaleidoscope of images, charts, and recordings to accompany their popular science lectures and scientific discoveries. It reveals Academia Sinica's real-time response to topics of concern in Taiwanese society and throughout the world.

Besides, this research is more dedicated to trying to reveal the establishment of websites and online curation. Because the expected readers are of a wide range of ages, it overcomes the “digital gap” and connects news and current affairs photos to create a refreshing design for the combination of science and humanities research. Designing simple and easy-to-understand click icons for a wider readership to create a novel and simple webpage operation mode is a point of this digital publishing project. In addition, in response to the needs of the public and with the increase in mobile viewing users, how to use RWD technology to provide mobile users with a new interface that is convenient for viewing in addition to the desktop browsing interface, making scientific research easier to popularize among people of all ages.

Therefore, how to use online curation in today's society where COVID-19 is raging, combining humanities and scientific research, combined with digital publishing and paper book publishing, to turn difficult knowledge into simple and easy-to-understand scientific text, and popularize it to the public through digital communication. The feedback from the public to observe the current social development under the epidemic is a new possibility of digital display that this research and this website are committed to trying.

---

<sup>1</sup> The Japan Foundation Research Fellowship, Visiting Researcher of Keio University Faculty of Letters, Academic Publishing Editor, Digital Museum Curator, and Research Fellow PhD of Academia Sinica Center for Digital Cultures. E-mail: wenjiunlinsera@gmail.com or as0200605@gate.sinica.edu.tw

## Introduction

The website "Reflections on COVID-19" is mainly based on the main policy of "Humanities Research and Digital Display" of Academia Sinica. It also responds to the establishment of the Academia Sinica Publishing Center and the implementation of the cross-disciplinary research project "Reflections on COVID-19". The main purpose and focus on the establishment of this research website are as follows:

- 1) International perspective and local observation: In response to important research-oriented topics in Taiwan and around the world, Academia Sinica gathers research scholars in related fields to join the project, and refer to the experience of international entity research and digital display to develop display interfaces and functional modules that adapt to the local area.
- 2) Cross-field cooperation and linkage: A relatively free communication platform is established with a topic-oriented approach. Scholars in various fields can speak independently from a professional perspective, or they can focus on discussions through collaboration, and accommodate a spectrum-like cooperation model. Interaction to organic integration.
- 3) Academia publishing and Online exhibition: Breaking through the traditional static and flat research publication form, fully recording the researcher's exploration history and archive materials, providing reader feedback and a mechanism for collaboration with the masses, and building a research exchange platform with the global academic community.

## Approach

### 1) Academia Sinica Publishing Center Preparation Plan and "Reflections on COVID-19" Digital Publishing Website Construction Plan

The development of information technology is constantly changing the methods of humanistic research. Since the Institute of History and Philology, Academia Sinica, started the digitization of Chinese ancient books in 1984 [1], for 37 years, "digital humanities" has continuously progressed and transformed from the collection, analysis and presentation of researcher data. The application of digital literature and computer networks has been an indispensable part [2].

Under this trend, in 2013, Academia Sinica Center for Digital Cultures (ASCDC) was officially established in the Institute of History and Philology, Academia Sinica [3]. On the one hand, it integrates the digital research talents and resources of our institute scattered in various research institutes. On the other hand, it continues to promote the "Digital Collection" and build a "Linked Knowledge Bases for Digital Humanities" that is suitable for humanities scholars. At the same time, ASCDC builds "digital humanities research tools" according to the needs of "humanities research", and establishes a open and transparent mechanism to recruit and manage the Academia Sinica's "Academic Research in the Digital Humanities" every year (Figure 1-1).

In May 2020, due to the global pandemic of COVID-19, our center (ASCDC) is committed to creating a platform that is more suitable for scholars to research and display

research results. As Academia Sinica considers digital publishing has been the trend of today's booming publishing industry, and we will prepare for the academic publishing center. The plan was entrusted to ASCDC to implement it. Since then, ASCDC planned and published the organization charter and business functions of the publishing center under the three ideal demands of "Digitalization", "Documentation" and "Demonstration" (Figure 1- 2), and even built the academic writing and digital publishing workflow (Figure 1-3).

From many perspectives, the major historical event in 2020 is the global disaster caused by COVID-19. This major event has not yet ended, and it still seriously affects the lives of human beings around the world. In the course of the development of the incident, how to use one's own knowledge to face the various changes brought about by the epidemic is the common concern of scholars regardless of discipline. The goal of natural science scholars is obviously relatively clear, which is to develop vaccines and seek ways that can alleviate or treat the condition.

"Reflections on COVID-19"[4] is a project of a group of humanities and sociologists at the Academia Sinica in the face of the 2020 epidemic. The researchers participating in the project use their own knowledge to try to understand the development of the epidemic and Related phenomena should be investigated, observed and understood, and the impact of the epidemic on knowledge and the relationship between different fields of knowledge should be explored. As the first important business of the Academia Sinica Publishing Center, the center is responsible for the two tasks of physical book and digital publishing, as well as the cultural promotion and marketing of the 2021 Taipei International Book Exhibition. The following is the "Reflections on COVID-19" website takes the construction as an example to discuss how "digital humanities" can use website construction more systematically and efficiently under the COVID-19 epidemic, build a bridge between scholars and the public, promote academics, and take care of intellectual responsibility to give back to the society (Figure 1-4).



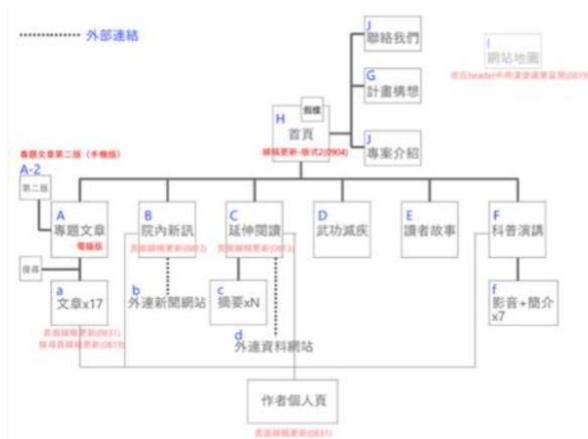
D) "Continuous updating": It is hoped that research and website display can keep pace with the times. The display (online exhibition) of research results is not an end point, but a starting point for conceiving new research. Moreover, through a feedback mechanism, researchers can reflect, revise, deepen, and expand new topics and methods.

Based on the requirements of the above-mentioned researchers, on the first-level homepage of the website structure, this website follows the basic website construction structure, planning project concept introduction, project introduction, and contact area. In order to make the page presentation more concise, the site map is changed to a "Hamburger Menu".

On the second level, we build and design six theme areas. "Feature Articles" contains all research articles of researchers. "News from the Academia Sinica" presents the research results related to COVID-19 in various institutes of the Academia Sinica with linked ways. "Extended Reading" is conducted through interviews with various scholars, and the researchers recommend reading chapters related to COVID-19. "Wugong(武功) Fights Diseases" combines ancient Chinese medical health exercises and digital mini games. "Reader Stories" provides a channel for website viewers' feedback and personal opinions. "Science Lectures" make it easier for website viewers to read audio-visual information related to COVID-19 lectures and programs of the Academia Sinica through audio-visual links (Figure 2-2).



2-1 : 「Reflections on COVID-19」  
Website Expected Goals



2-2 : 「Reflections on COVID-19」  
Website Basic Function Construction (User Interface Design)

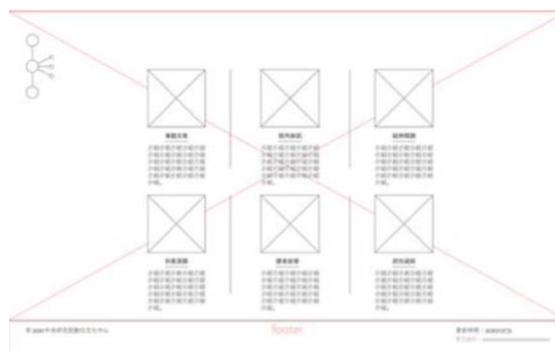
**Figure 2: "Reflections on COVID-19" Website Basic Function Construction Plan and Expected Goals**

**3) "Reflections on COVID-19" Website Art Editing and Design Features**

Regarding the art editing and design of the website of "Reflections on COVID-19", due to the limited funding at the beginning of the publishing center's establishment, the website design adopts an intuitive and concise method. In terms of text, it is selected and recommended by each researcher, and they select the "key sentences" and "summary content" that are suitable to represent the research purpose of the researcher (Figure 3-1). On the other hand, all web design adopts news and current affairs photos, or scholars research material images, because using bright photos or the picture reminds viewers of contemporary historical events, and visually brings the viewer the impression that history and society are closely related (Figure 3-2). At the same time, the use of this design can eliminate external manufacturer's web design costs, effectively saving costs, and in terms of legal authorization of website pictures, it can simplify the authorization signing process and ensure that all pictures comply with laws and regulations and can be legally disclosed.



3-1 : "Reflections on COVID-19" website mainly page on UI design



3-2 : Theme planning and UI design of each unit of the "Reflections on COVID-19" website

**Figure 3: "Reflections on COVID-19" Website Art Editing and Design Features**

#### 4) "Reflections on COVID-19" Website is Publicly Launched and Web Page

##### Introduction

Starting from the "Reflections on COVID-19" website construction plan in July 2020, the website has been launched in April 2021. The planning and design of the website functions are subject to changes and adjustments in response to various requests from researchers. The following are explanations on the published webpage of "Reflections on COVID-19."

First of all, the main color of the website design is "black". Take Figure 4-1 as an example. The black design is to remind the readers of the severity of the COVID-19 pandemic and bring about the feeling of caution and fear during the pandemic. At the same time, it can also improve the color contrast between the pictures, set off the photo or research picture selected by the researcher, and the visual design of the main screen, each

page is presented with photos related to the research, as shown in the background of Figure 4-1. The picture is a photo of " the Burning of the King Boat(燒王船)" traditional ceremony used by the people in the Donggang (東港) area of southern Taiwan to eradicate the disease in a study conducted by our researcher Paul R. Katz. Figure 4-2 is the extended reading material recommended by the researcher. As COVID-19 affects not only the health and psychology of the public, but also includes the National Taiwan University Entrance Examination, subject examinations are also adjusted in line with current events. Scientific test questions appeared about COVID-19-related virus medical knowledge, and Mandarin and English reading tests appeared about COVID-19-related internal and external reports. Therefore, the provision of extended reading materials can satisfy the viewer's desire for knowledge, and at the same time, it can also achieve the function of assisting high school and middle school education.

Figure 4-3 presents the eight important parts of this website in the form of easy-to-understand picture icons, namely "Feature Articles", "Information of Academia Sinica", "Extension Reading", "Science Lectures", "Readers' Stories", and "Wu Gong(武功) Fight Diseases", "Conception of Plan", "Contact Us". After clicking on each picture icon, you can simply and intuitively use the link to obtain more relevant information. Figure 4-4 is the " Science Lectures" link to COVID-19 related speeches lecture activity.



4-1 : Introduction to the Main Page Design of "Reflections on COVID-19"



4-2 : "Reflections on COVID-19" Website Extension Reading Page



4-3 : Introduction to the Themes of Each Unit of the "Reflections on COVID-19" Website



4-4 : Introduction to the Link Page of Academia Sinica's Activities on the "Reflections on COVID-19" website

## **Figure 4: "Reflections on COVID-19" Website is Publicly Launched and Web Page Introduction**

### **5) Responsive Web Design (RWD) to Meet the Needs of Mobile and Tablet Users**

In response to the rapid development of technology and the changes of the times, nearly 80% of modern people are accustomed to using mobile phones to search, browse websites, and shop. Coupled with the popularization of tablets, more and more screens of different sizes have appeared, resulting in the original the problem that the webpage cannot be displayed normally. In the previous situation of the old version of the website, you may have encountered a different layout ratio on the computer and the mobile phone when browsing some corporate websites or shopping websites. It may be that the entire layout is cropped and the picture cannot be seen. Or the pictures and fonts are very small, making the operation very uncomfortable, resulting in the direct exit of the page at the end. Therefore, this research website adopts responsive web design (RWD), that is, the page will be automatically adjusted to the most suitable page depending on the mobile device you use.

Responsive Web Design is abbreviated as RWD, which is a concept proposed by Ethan Marcotte, a top foreign designer, started in 2011. After 2012, it is recognized as the trend of web design and development technology in the future, and its principle is to give instructions through grammar, use HTML5, CSS3 and other programming languages. The actual operation mode will set interrupts for different sizes of web pages, and divide the size of the web page into many sections from large to small, and each section will set a suitable layout. As the size of the browser changes, the web page will apply different CSS settings, so that the layout will be adjusted accordingly, and that the website can self-adaptation with perfectly match the size of the device and provide users the best visual experience. It is as if water is packed in a different container, it will become a concept of a different shape.

In particular, in November 2014, the global search engine leader Google's search algorithm experienced a major change. The new search results will allow those sites with good experience on mobile devices to have the advantage of prioritization, and the responsive website design is the site with a good experience in Google's review. RWD websites are suitable for viewing on different device platforms, such as mobile phones, tablets, laptops, desktops...etc., they are all suitable, and can automatically present suitable sizes and web page styles. Easily improve website ranking and increase the number of visitors. On the whole, making all pages of the website have a responsive design, that is, allowing users to clearly see all web content on devices of any screen size, which is one of the important factors in attracting Google's natural search traffic more effectively.

In the process of the research website development and construction, the advantages and disadvantages of the RWD website were sorted and summarized as follows.

### **A. The advantages of RWD responsive web pages**

A-1. The development cost is low, and it solves the problem of browsing on multiple devices. Only the cost of making a website is required. At the same time, the production period is relatively short, which meets the requirements of website design that needs to be completed in a short time.

A-2. It is convenient to maintain the content of the website, only one website needs to be managed, and the cost of manpower management is also saved simultaneously.

A-3. To improve the effectiveness of website marketing, duplicate content is a big problem for search engines, which easily affects the ranking of web pages on search engines. The use of RWD web design can completely avoid duplicate content due to version differences.

A-4. Supports cross-platform devices, which can be used on any device, and it is much more convenient to share webpages. When users share a website, sometimes they may not get good browsing results because other users open the website on different devices. Because of RWD is only for a website, there is only one sharing URL, so these problems will not occur.

A-5. The visual image is consistent, which enhances the viewer's impression of the website. You don't have to be busy zooming in. The convenience of the user's browsing will increase the user's stay time and increase the impact of the webpage on the viewer and the impression in the brain.

### **B. Disadvantages of RWD responsive web pages**

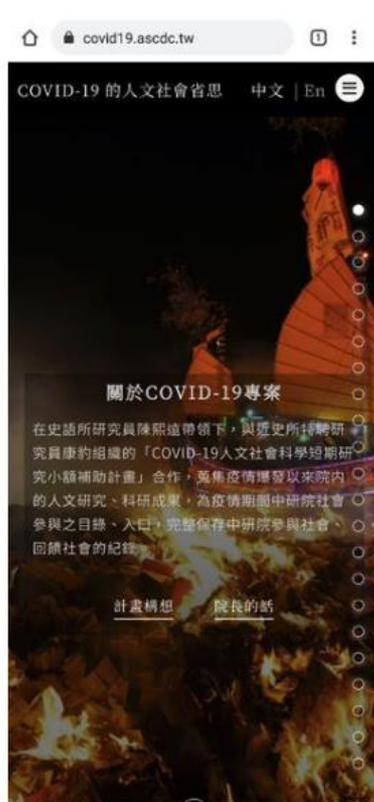
B-1. The data capacity of the website may be large, which affects the slower and longer loading speed of the website.

B-2. Compared with the simple mobile website, the content loaded by RWD is the full computer version of the website. The extra content information is hidden when browsing on the mobile phone. Therefore, the website takes a long time to load and the mobile user feels it is even more obvious. But this problem has been slowly solved with the efforts of developers to optimize and the speed of mobile devices and websites. Because multiple devices share web pages, it is not suitable for more complex functional interfaces. This is a common disadvantage of all small screens. Due to the limited screen space and the inherent limitations of mobile devices with

only touch functions, it is not easy to use RWD to develop some applications on the "desktop or laptop" web page function.

B-3. Older browsers are not supported-because RWD uses the latest HTML5 with CSS3 web technology to process, some older browsers may not be compatible. If you use an outdated browser to watch, the layout may be broken. However, this problem has been liberated with Microsoft's abandonment of versions below IE10. Most browsers now support HTML5. If there are incompatibility issues, users can be advised to upgrade and update their browsers.

B-4. The current production method of RWD webpages still cooperates with webpage manufacturers to cut the browsing screen. Therefore, when designing the user interface, it is necessary to redraw the wireframe draft suitable for mobile phones. Moreover, if the visual presentation of each page of the screen were large, which must be revised and adjusted repeatedly in the user's demand design planning and website interface confirmation.



5-1 : Mobile version main page design



5-2 : Each Unit Screen of the Mobile Version



§The above images are all 6.5-inch 21:9 Cinema Wide 4K HDR OLED screen viewing effects

## **Figure 5: RWD Responsive Web Design to Meet the Needs of Mobile and Tablet Users**

### **6) Ideal and Reality: Cultural Dissemination and Digital Marketing Promotion under COVID-19**

This research website was originally expected to be set up at the Taipei International Book Fair in 2021 in the exhibition area of the Taipei International World Trade Center. It will use digital devices, AR, and VR to conduct "immersive experience" digital reading activities, allowing visitors to see through the large-scale exhibition screens, to experience the sensory stimulation brought by the digital display.

However, due to the spread of COVID-19, the 2021 Taipei International Book Fair had to stop the exhibition. ASCDC chose to host and transform the originally planned exhibition into an online exhibition [5]. On January 26, 2021, we launched "Advancing Knowledge! Academia Sinica Special Exhibition Event Website", linked our precious collections and publications, digital platform and database, so that the public could still be online to have a further understanding of the creation and evolution of knowledge in the digital age, and cooperate with this research website to provide different visual experiences and knowledge dissemination.

In addition, during the exhibition, ASCDC hosted "Digital Publishing and Academic Publishing of the COVID-19 Humanities and Social Reflection Project" speech, and conducted an online live broadcast event on Facebook [6], allowing readers to directly communicate with the speaker and host of the live speech through the live broadcast. People interact to learn more about the popular science function played by the "Reflections on COVID-19" website. Besides, ASCDC continues to update the "Reflections on COVID-19" related information link on the official Facebook website to continuously respond to the public's need for knowledge about current affairs issues and COVID-19 related research. At the same time, humanities research is combined with digital technology to provide various new experiences of "discussing old learning" and "cultivating new knowledge" online.



6-1 : 2021 Taipei International Book Fair Exhibition Area and Website Plan of Academia Sinica

6-2 : 2021 Taipei International Book Fair Academia Sinica Online Exhibition (actual status)

**Figure 6: Ideal and Reality: Cultural Dissemination and Digital Marketing Promotion under COVID-19**

**7) Combination of ancient Chinese medical history and digital humanistic interest**

Life is a path of cultivation. The concept of cultivation does not detract from its value due to technological civilization. On the contrary, it can give renewal meaning and guide the direction of human development.

Due to the limitations of the human perception system, there are many people who can detect outwards and fewer can detect them inwardly. Most people understand that their body is only limited to the concept of anatomical or molecular biological parts. When it comes to health preservation, only balanced nutrition and moderate exercise are known. The rest is nothing. Never listened to the sound of the body and communicated with it. "Vipassana" is what Chinese medicine emphasizes. For more than a hundred years, we do not cherish our own cultural characteristics. Although Han people do not necessarily wear Tang suits, cheongsams, or sit on Chinese-style seats, they must face up to and appreciate the value and significance of traditional culture. The next choice is to go to the gym to do gravity, aerobic training, or play Ba Duan Jin (八段錦), practice calligraphy, or learn Chinese painting, which is more beneficial to the body and mind.

Recently, the COVID-19 pneumonia epidemic has been serious and tightened, and "how to improve resistance" has become a topic of concern to everyone. From the perspective of Chinese medicine, if you can practice the "Ba Duan Jin (八段錦)" of Chinese medicine, it will help relieve muscles and bones, relieve muscle aches, and enhance your own immune system. It is a good home exercise choice for those who cannot go out for exercising.

This website uses digital technology combined with traditional Chinese health Knowledge. The Ba Duan Jin (八段錦) atlas, which has been collected in the National Palace Museum( in Taipei now) since the Qing Dynasty, is designed as a graphic card and

made into an online game, which can be viewed in a way of "education and fun". Through simple and interesting online games, people learn about the theories of Traditional Chinese medicine health-preserving viscera, meridians, acupuncture points, muscles, meridians, qi, blood and body fluid knowledge, which are close to traditional culture. Coincidentally, during the epidemic of COVID-19, people can use online games to release their pressure and enhance the pleasure of daily life.



7-1 : The "Ba Duan Jin(八段錦)" Health Exercise in the Qing Dynasty of China



7-2 : Card Game Design and "Ba Duan Jin(八段錦)"

**Figure 7: Combination of Ancient Chinese Medical History and Digital Humanistic Interest**

## Conclusion

This research attempts to explore new possibilities for digital display with the website "Reflections on COVID-19". Due to the COVID-19 epidemic, the website has been postponed to be publicized. Therefore, this website is currently being continuously updated. In the future, in addition to the function of timeline, digital history maps on the website, we will also try to collect readers' online feedback content, to extract keywords from it, perform text analysis and add text mining and other functions. We hope to use digital tools and the digital display platform, to record the major events of the times, analyze the impact of disease on the world, and try to expand the new horizons of contemporary history and digital humanities research.

## Reference

- [1]. Scripta Sinica database, <http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm>, Accessed on 2021-08-18.
- [2]. Taiwan Digitalarchives, <https://culture.teldap.tw/culture/index.php>, Accessed on 2021-08-18.
- [3]. Academia Sinica Center for Digital Cultures, <https://ascdc.sinica.edu.tw/en/>, Accessed on 2021-08-18.

- [4]. Reflections on COVID-19, [https:// https://covid19.ascdc.tw/en](https://covid19.ascdc.tw/en), Accessed on 2021-08-18.( <https://covid19.ascdc.tw/>)
- [5]. Advancing Knowledge! Academia Sinica Special Exhibition Event Website, [https:// https://evoread.ascdc.sinica.edu.tw/](https://evoread.ascdc.sinica.edu.tw/), Accessed on 2021-08-18.
- [6]. Digital Publishing and Academic Publishing of the COVID-19 Humanities and Social Reflection Project, [https:// https://zh-tw.facebook.com/ASCDCNEWS/videos/2806614029589412/](https://zh-tw.facebook.com/ASCDCNEWS/videos/2806614029589412/), Accessed on 2021-08-18.

# **Sonifying the pandemic – innovative approaches towards data interaction and engagement formats for scientific, educational and artistic purposes**

**Michael Stark<sup>1</sup>, Amelie Dorn<sup>2</sup>, Renato Rocha Souza<sup>3</sup>**

## **Introduction**

This paper introduces an innovative digital sonification prototype based on openly available COVID-19 data for Austria, Europe. Different sinusoidal waves, correspond to the nine counties of Austria (Vienna, Burgenland, Upper Austria, Lower Austria, Salzburg, Carinthia, Tyrol, Styria, Vorarlberg) were set up; the amplitude of the sound generators was modulated by the daily cases of COVID-19-infections, and changes in pitch were achieved by mapping the 7-day-incidence according to the traffic-light warning system established by the Austrian government. Besides scientific, artistic and informational purposes, the prototype can serve as a valuable, innovative and interactive educational tool.

## **Approach**

The arrival of the Coronavirus pandemic in 2020 has brought unprecedented effects to all countries and populations worldwide. As COVID-19 developments continue, countries are evolving their strategies, realizing that the pandemic's risk is multidimensional with long term consequences. Apart from its apparent threat to human health, the World Health Organisation (WHO) has also attested that the COVID-19 outbreak is accompanied by an “infodemic” [1]. Although this is not a novelty tied to COVID-19, the sheer amount of mis- and disinformation presents a threat to public health and public action [2]. One possible explanation is the lack of clarity, described as epistemic uncertainty [3]. While experts are learning in real-time about the phenomenon, it is clear that knowledge about it will change over time [4]. Another factor to be taken into consideration is related to the various quantitative measurements of COVID related dimensions. Many measured parameters, like the case-fatality-rate, the total number of infected people, or accuracy in testing, can be distorted, making it hard to consistently convey coherent facts [5].

A methodology that has proven to be useful at providing support for understanding great amounts of heterogeneous data — especially when several dimensions of data need to be monitored simultaneously — is data sonification [6]. Interdisciplinary by nature, data sonification aims at making sounds auto-explanatory, extrapolating its pure

---

<sup>1</sup> University of Vienna

<sup>2</sup> Austrian Centre for Digital Humanities and Cultural Heritage, Austrian Academy of Sciences (ACDH-CH OeAW)

<sup>3</sup> Centre for Image Science, Danube University Krems

musical context and conveying more information than traditional data formats [7]. As historic examples such as the Geiger-Counter, the Pulse-Oximeter or several data analysis and exploration tasks related to the Voyager 2 mission have shown, sonification systems have always been helpful in displaying structural and temporal aspects of complex signals that would not have been as efficient with visual tools alone [6]. This synergetic support for understanding data derives not only from its accessibility through interactivity, but also from some of the basic features of the human auditory system such as its ability to handle multiple channels of sound at the same time, being sensitive to rhythmic patterns or its capacity of perceiving variations in sound with high temporal resolution [8].

### Application Design

Based on these observations, the aim of this work was to create an interactive data sonification prototype which was coded using the Python programming language. The current prototype aims at functioning as a springboard for the development of an interactive digital data sonification museum, aiming to serve as a valuable, innovative and interactive educational device, besides scientific, artistic and informational purposes. We hold the assumption that sonification of data sets can evoke more learner engagement especially in younger students. Historically it has been demonstrated, from a didactics perspective, that the use of rhythm and melodies can positively influence learning processes, e.g. in teaching the alphabet, or the number of days in each month [6].

For the design of the prototype [9] that sonifies the current open source data on COVID-19 for Austria [10], the use of traditional instruments, sounds or musical motifs was avoided. This was done for several reasons: musical motifs and sounds are culturally shaped and, therefore, lack universal meaning [11]. For these reasons, sine-tones were chosen as the main building blocks. Due to the same considerations, musical motifs were substituted by multiplication factors of base frequencies according to a mapping system describing pandemic risk levels. This also avoids stepping into the more complex territory of using Musical Instrument Digital Interfaces (MIDI), since MIDI-information also corresponds to formal musical notes [12].

As a first step in the data-to-sound mapping, nine sinusoids corresponding to the nine counties of Austria (Vienna, Burgenland, Upper Austria, Lower Austria, Salzburg, Carinthia, Tyrol, Styria, Vorarlberg) were set up. The base frequency of each sine-tone correlates with the latitude-value of the county in a way that the latitude values are sorted and then assigned to 9 different frequencies. These were derived from the multiplication of a base frequency with one of three sets of multiplication-factors [neutral<sup>4</sup>, light<sup>5</sup>, dark<sup>6</sup>].

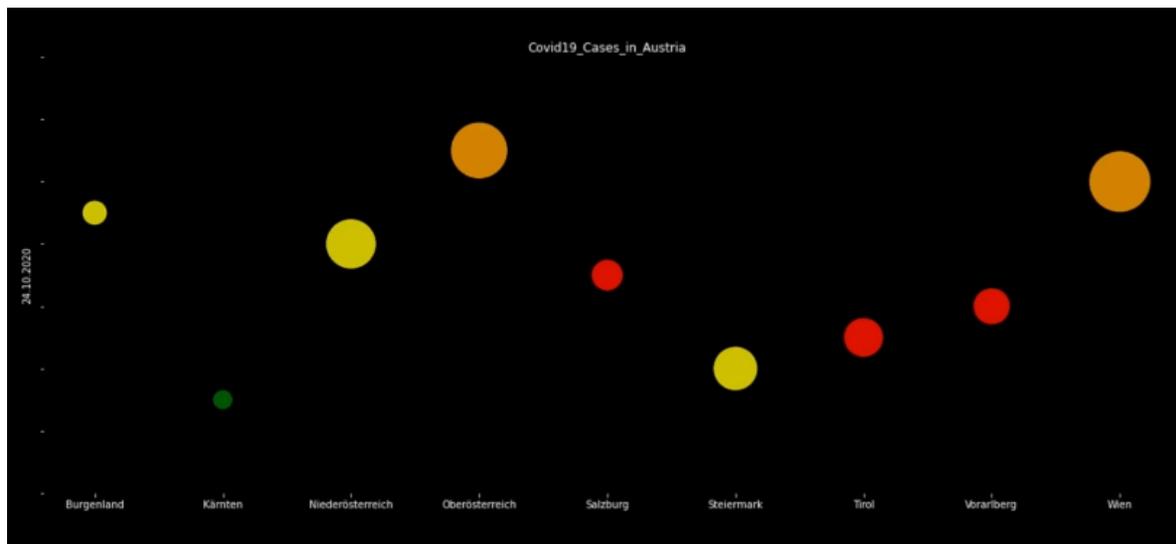
---

4 [\*1, \*1.5, \*2, \*3, \*(4 \* (9/8)), \*5, \*(4 \* (15/8)), \*8, \*9]

5 [\*1, \*1.5, \*2.5, \*3, \*3.75, \*4.5, \*5.625, \*6, \*7.5]

6 [\*1, \*2, \*2.4, \*3, \*3.5, \*4.5, \*4.75, \*6, \*7]

Both aspects provide a point of interaction for recipients allowing to shape the narrative of the sonification drastically. For the prototype a frequency of 110 Hz and the “neutral” set were chosen. The placement of the nine sine-tones in the stereo field corresponds to the longitude-values of the nine counties. The amplitude of the sound generators is modulated by the daily cases of COVID-19-infections, where the numbers were scaled to represent a value between 0 and 1, to suit the needs of being a relevant amplitude value. Pitch changes were achieved by mapping the 7-day-incidence according to the aforementioned warning system, established by the Austrian government. The warning system that corresponds to a traffic light was used to make government actions tangible, relying heavily on the 7-day-incidence-values [13]. The mapping was handled in a way that each step further towards the red light, the base frequencies get multiplied by a given number<sup>7</sup>, increasing in pitch. This resulted in a dense, loud and high-pitched cluster at the peak times of the pandemic, and lower pitches otherwise (Figure 1).



**Figure 1: Visualisation of the animated sonification scatter-plot with the nine Austrian counties (x-axis) and time (y-axis).**

As a next step, further parameters will be introduced via frequency modulation, which allow for enhanced complexity in sound without losing its commitment to sine waves as the main building blocks.

## References

- [1]. World Health Organisation (WHO). (2020). Novel Coronavirus (2019-nCoV) Situation Report-13. Online article. <https://www.who.int/docs/default->

<sup>7</sup> [\*1, \*1.5, \*2, \*2.5, \*3]

- source/coronaviruse/situation-reports/20200202-sitrep-13-ncov-v3.pdf [last access: 17.08.2021]
- [2]. Brennen, S., Simon, F., Howard, P. N. & Kleis Nielsen, R. (2020). Types, sources, and claims of COVID-19 misinformation. Online article. <https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation> [last access: 17.08.2021]
- [3]. Rocha Souza, R.; Dorn, A.; Piringer, B.; Wandl-Vogt, E. (2019). Towards a Taxonomy of Uncertainties: Analysing Sources of Spatio-Temporal Uncertainty on the Example of Non-Standard German Corpora. *Informatics*, 6, 34. <https://doi.org/10.3390/informatics6030034>
- [4]. Caplan, Robyn (2020). COVID-19 misinformation is a crisis of content mediation. Brookings. Online article. <https://www.brookings.edu/techstream/covid-19-misinformation-is-a-crisis-of-content-mediation/> [last access: 17.08.2021]
- [5]. Lachmann, A. (2020). Correcting under-reported COVID-19 case numbers. MedRxiv.
- [6]. Kramer, G., Walker, B., Bonebright, T., Cook, P., Flowers, J. H., Miner, N., & Neuhoff, J. (2010). Sonification report: Status of the field and research agenda.
- [7]. Grond, F., & Hermann, T. (2012). Aesthetic strategies in sonification. *AI & Soc* 27, 213–222.
- [8]. Hermann T., & Ritter, H. (2004). Sound and Meaning in Auditory Data Display. *Proceedings of the IEEE (Special Issue on Engineering and Music - Supervisory Control and Auditory Communication)* 92(4), 730-741.
- [9]. [Sonification Prototype [https://michaelxstark.github.io/Data\\_Son\\_COV-19\\_AUT\\_2021/](https://michaelxstark.github.io/Data_Son_COV-19_AUT_2021/)] [last access: 17.08.2021]
- [10]. Open Data Österreich <://www.data.gv.at/> [last access: 18.08.2021]
- [11]. Vickers, P., & Hogg, B. (2006). Sonification Abstraite/Sonification Concrète: An 'Aesthetic Perspective Space' for Classifying Auditory Displays in the Ars Musica Domain. Paper presented at ICAD 2006 - the 12th Meeting of the International Conference on Auditory Display, UK June 20 - 23.
- [12]. Smith III, Julius O. (2005). Viewpoints on the History of Digital Synthesis. Portability and the Limits of MIDI. *Proceedings of the International Computer Music Conference (ICMC-91, Montreal)*, pp. 1-10. [https://ccrma.stanford.edu/~jos/kna/Portability\\_Limits\\_MIDI.html](https://ccrma.stanford.edu/~jos/kna/Portability_Limits_MIDI.html) [last access: 17.08.2021]
- [13]. Corona Ampel Österreich. Website. <https://corona-ampel.gv.at/corona-kommission/bewertungskriterien/> [last access: 17.08.2021]

# Thailand Towards Digitization– the past, the present, the future and gray digital gap

Saiyud Moolphate<sup>1</sup>, Nadila Mulati<sup>2</sup>, Thin Nyein Nyein Aung<sup>2</sup>, Motoyuki Yuasa<sup>2,3</sup>,  
Myo Nyein Aung<sup>3,4</sup>

## The past

The digitalization of Thailand started in February 1996, the first National IT policy (IT2000) was launched (1996-2000) that was aimed to integrate ICT into government administrations, increase ICT capacity and invest in national information infrastructures. This is the first information and technology policy that had been launched. After the successful framework was provided to policies and projects, Thailand launched ICT Policy Framework (IT2010) to further develop ICT. IT2010 is a ten-year plan, from 2001 to 2010, focused the development of long-term “5-Es Strategy”, which were e-Government, e-Industry, e-Commerce, e-Education and finally e-Society[1]. Under this framework, the first National ICT Master Plan (2002-2006) and the second National ICT Master Plan (2009-2013) were launched to further execute and realize the IT2010. Thailand 4.0 initiative and the Digital Thailand plan were launched in 2016[1]. The ASEAN ICT Masterplan 2020 (AIM 2020) was endorsed by ASEAN ICT Ministers at the ASEAN Telecommunications and Information Technology Ministers Meeting (ASEAN TELMIN) in November 2015[1]. 20-Year National Strategy (2018-2037) was the first national long-term strategy developed according to the Constitution of the Kingdom of Thailand 2017[3] [4].

## The Present

A rural-urban ‘digital divide’ persists and a 2016 national survey finds that while more than 70 percent of the Bangkok urban population uses smartphones, only 39 percent do in the more rural and less developed Northeastern region[5]. The number of households connecting to the internet classified by regions in 2017 was 4.6 million in the central region, whereas 2.1 million households in the northern region. Internet usage of people who are not in the labor force (kids and elderly);158,264 people (3.36 percent) and that of patients, and disabled Persons was 128,163 people (9.21 percent) by the National

---

<sup>1</sup> Department of Public Health, Faculty of Science and Technology, Chiang Mai Rajabhat University, Chiangmai 50300, Thailand

<sup>2</sup> Department of Public Health, Graduate School of Medicine, Juntendo University, Tokyo 113-8421, Japan

<sup>3</sup> Faculty of International Liberal Arts, Juntendo University, Tokyo 113-8421, Japan

<sup>4</sup> Advanced Research Institute for Health Sciences, Juntendo University, Tokyo 113-8421, Japan

Statistical Office Thailand in 2017. Thai people's digital literacy was at a basic level with a score of 64.5 (out of 100) in 2018[6].

The following factors determined the usage of the internet by the elder people; 1) Education status of bachelor degree and above, 2) Residences in the urban, compared to rural area, 3) Male, 4) Retired, 5) Having experience of usage of computer/smart phone, tablet, 6) Staying in the area having internet access, 7) Co residence with people using internet[7] [7] [7].

### **The Future**

The COVID-19 global pandemic accelerated the digitalization of the countries and also highlighted the importance of ICT to achieve Sustainable Development Goals (SDGs) [6] Therefore, seniors must be able to use digital technology effectively. Governments and public-private sectors should all work together to establish policies and strategies to empower the Thai senior population, realize "leave no one behind".

### **Reference**

- [1]. Heeks, R. and Bukht, R. (2018). Digital Economy Policy: The Case Example of Thailand. SSRN Electronic Journal doi:10.2139/ssrn.3540030.  
<https://www.ssrn.com/abstract=3540030> (accessed 9 August 2021).
- [2]. Bolliger & Company (Thailand) Ltd. (2016). Roadmap for Thailand's Implementation under the ASEAN ICT Masterplan 2020; Office of the Permanent Secretary Ministry of Digital Economy and Society: 2016; p 120  
[https://www.onde.go.th/assets/portals//files/AIM\\_2020%20\\_Thai\\_Eng.pdf](https://www.onde.go.th/assets/portals//files/AIM_2020%20_Thai_Eng.pdf).
- [3]. Ministry of Information and Communication Technology Digital Thailand: Thailand digital economy and society development plan  
[https://www.onde.go.th/assets/portals/files/Digital\\_Thailand\\_pocket\\_book\\_EN.pdf](https://www.onde.go.th/assets/portals/files/Digital_Thailand_pocket_book_EN.pdf).
- [4]. National Strategy Secretariat Office Office of the National Economic and Social Development Board (2018). NATIONAL STRATEGY 2018 – 2037.  
<https://sto.go.th/sites/default/files/2019-12/National-Strategy-Eng-Final-25-OCT-2019.pdf>.
- [5]. National Statistical Office and Ministry of Digital Economy and Society (2016). The 2016 Household Survey on the Use of Information and Communication Technology. Annual Survey.
- [6]. Office of the National Digital Economy and Society Commission (2019). The state of Thailand's digital economy and society development 2019; Office of the National Digital Economy and Society Commission  
[https://www.onde.go.th/assets/portals//files/Booklet\\_1.pdf](https://www.onde.go.th/assets/portals//files/Booklet_1.pdf).

- [7]. National Statistics Office and Ministry of Information and Communication Technology (2015). Analytical Report: Factors Affecting Internet Use of Elderly in Thailand.
- [8]. Loipha, S. (2014). Thai Elderly Behavior of Internet Use. *Procedia - Social and Behavioral Sciences*, 147 doi:10.1016/j.sbspro.2014.07.125.
- [9]. Electronic Transactions Development Agency, Ministry of Information and Communication, and Technology (2016). Thailand Internet User Profile 2015. [https://unctad.org/meetings/en/Contribution/dtl\\_eweek2016\\_ETDA\\_IUP\\_en.pdf](https://unctad.org/meetings/en/Contribution/dtl_eweek2016_ETDA_IUP_en.pdf).
- [10]. Johnson, M. (2021). Technology and older persons: Ageing in the digital era. International Telecommunication Union <https://www.itu.int/en/myitu/News/2021/02/08/17/18/Technology-older-persons-ageing-digital-era-Malcolm-Johnson>.

# Wikidata as a Low-tech Solution to Leverage Semantic Technologies and A Case Study of CBDB ID's Reconciliation with Wikidata

Fudie Zhao<sup>1</sup>

While the importance of the Semantic Web technologies has been widely acknowledged among the DH community, the difficulties in acquiring relevant technical skills may hamper many DH practitioners. Wikidata may be one of the low-tech solutions to leverage semantic technologies. This short paper will present a case study of evaluating the benefits and problems in the reconciliation of CBDB ID with Wikidata to demonstrate this point. By doing so, it intends to invite discussion from participants at JADH regarding Wikidata as a Linked Data approach to integrate data and as an entry to disseminate data to the Web in East Asia's context and the challenges it may face.

## What is Wikidata

The Wikidata repository consists mainly of items. As a sister project of Wikipedia, it maintains a simple, user-friendly, collaboration-oriented interface for editing items.<sup>2</sup> On its user interface, an item is displayed in 4 parts: **1) basic information** includes label, item identifier, description, aliases, and their multilingual displays. **2) statements**, which describe features of an item, consists of property and value. **3) external identifiers** link an item to external databases. **4) sitelinks** connect an item to other Wikimedia projects, like Wikipedia.

Behind its simple user interface are interlinked item and their data supported by semantic technologies. Featuring a user-generated ontology,<sup>3</sup> a live SPARQL endpoint for data query,<sup>4</sup> regular RDF dumps,<sup>5</sup> and interlinks to other open datasets,<sup>6</sup> it is Wikimedia's response to semantic technologies. It claims that following its tutorials, even novice users can create and publish semantically-rich structured data conforming to Wikidata's ontology.<sup>7</sup> Its strong community has developed various applications and tools to support data publication and integration which further lower the technical barrier.<sup>8</sup>

---

<sup>1</sup> The University of Oxford (fudie.zhao@sant.ox.ac.uk)

<sup>2</sup> See the English page about Confucius on Wikidata as an example:  
<https://www.wikidata.org/wiki/Q4604>

<sup>3</sup> [https://www.wikidata.org/wiki/Wikidata:WikiProject\\_Ontology](https://www.wikidata.org/wiki/Wikidata:WikiProject_Ontology)

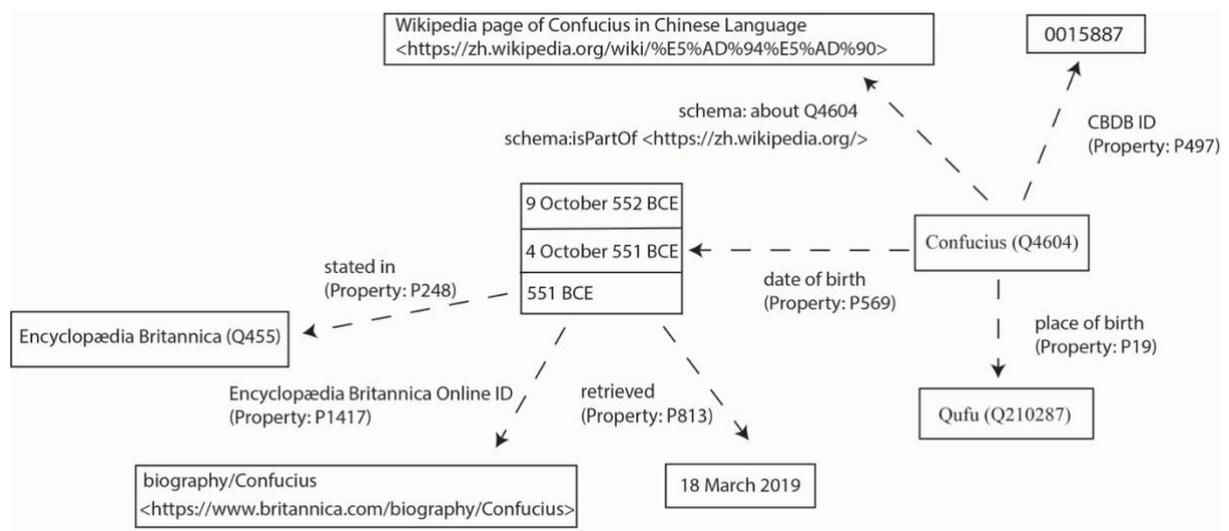
<sup>4</sup> <https://query.wikidata.org>

<sup>5</sup> [https://www.wikidata.org/wiki/Wikidata:Database\\_download](https://www.wikidata.org/wiki/Wikidata:Database_download)

<sup>6</sup> <https://tools.wmflabs.org/mix-n-match/#/>

<sup>7</sup> <https://www.wikidata.org/wiki/Wikidata:Tours>

<sup>8</sup> <https://www.wikidata.org/wiki/Wikidata:Tools>



**Figure 1: Linked data behind the user-friendly interface of Wikidata (the English page about Confucius as an example).**

### What benefits has the integration brought to CBDB?

Although Wikidata is not mentioned on CBDB's official sites, 421,006 items on Wikidata had a CBDB ID by 16 August 2021.<sup>9</sup> One way to measure the effectiveness of Wikidata-facilitated data integration and dissemination for CBDB is to examine the number of **external identifiers** and **sidelinks** that items with CBDB ID have on Wikidata, as these external identifiers and sidelinks are pointing to data in other systems. Information about external identifiers can be queried via the SPARQL portal on Wikidata. For an overview, these items are linked with 361 external identifiers other than CBDB person ID.<sup>10</sup> The query results show that linking to Wikidata has:

- 1) enhanced integration in the historical Chinese domain by collaboration between DH projects and GLAM. For example, the Wikidata item of Confucius has 127 **external identifiers**; most are from GLAM domains.<sup>11</sup> By linking to Wikidata,

<sup>9</sup> Data on Wikidata is constantly changing. Wikidata data provided here are query results on 16 August 2021. URLs will be provided in this paper for readers to access the SPARQL live query results. Check how many Wikidata items have a CBDB ID when you are reading this paper via the following link: <https://w.wiki/3NYY>

<sup>10</sup> <https://w.wiki/3NZ3>

<sup>11</sup> <https://w.wiki/3NZc>

datasets from these institutions and CBDB can be mutually contextualised for the information concerning Confucius.

- 2) increased the CBDB data's visibility to the general public. Wikidata is interlinked with Wikipedia which is more widely known by the public. **Sitelinks** between Wikimedia's projects can measure interlinks between an item in Wikidata and its corresponding Wikipedia page.<sup>12</sup> So far, 12,903 items with a CBDB ID in Wikidata have been linked with corresponding pages in Wikipedia in the Chinese language.<sup>13</sup> Displaying the CBDB ID in Wikipedia, however, still requires extra effort.<sup>14</sup> The publication of CBDB ID on Wikidata can be taken as the first step to disseminate its data to Wikipedia.

### **What are the problems and potential solutions?**

The reconciliation has been conducted by the Wikidata community without any collaboration with CBDB. This independent development of CBDB and Wikidata so far has brought at least two problems to be solved: 1) How to locate and correct identifier mismatches caused by the lack of inspection during the data export? 2) Both CBDB and Wikidata have evolved after the reconciliation occurred. How to take this ambivalent and changing nature into account for sustainable integration in the future? It is beyond this paper's scope for these problems to be thoroughly addressed. Instead, here we will focus on some possible solutions that Wikidata may offer.

For problem one, automated inspection is customisable by setting up **property constraints** for an **external identifier** on Wikidata. For example, **single value constraint**, which specifies that a property generally has only a single value, can detect Wikidata items that have more than 1 CBDB ID. Items that violate this constraint are displayed on the **constraint violation report** for CBDB ID.<sup>15</sup> Wikidata provides many types of **property constraint**, and with some knowledge of SPARQL, users can customise complex constraints for a property.<sup>16</sup>

For problem two, Wikidata does not take the responsibility to keep the data up to date. Instead, it encourages data providers to maintain the data either by themselves or by preparing documentation for the Wikidata community to help with it.<sup>17</sup> Wikidata has features that keep data changes trackable. Changes to items related to CBDB ID have been

<sup>12</sup> <https://www.wikidata.org/wiki/Help:Sitelinks>

<sup>13</sup> <https://w.wiki/3TnK>

<sup>14</sup> See the section "Inserting Wikidata values into Wikipedia articles" in [https://en.wikipedia.org/wiki/Wikipedia:Wikidata#Inserting\\_Wikidata\\_values\\_into\\_Wikipedia\\_articles](https://en.wikipedia.org/wiki/Wikipedia:Wikidata#Inserting_Wikidata_values_into_Wikipedia_articles)

<sup>15</sup> [https://www.wikidata.org/wiki/Wikidata:Database\\_reports/Constraint\\_violations/P497](https://www.wikidata.org/wiki/Wikidata:Database_reports/Constraint_violations/P497)

<sup>16</sup> [https://www.wikidata.org/wiki/Help:Property\\_constraints\\_portal#Usage\\_instructions](https://www.wikidata.org/wiki/Help:Property_constraints_portal#Usage_instructions)

<sup>17</sup> [https://www.wikidata.org/wiki/Wikidata:Data\\_donation#4.\\_Keep\\_the\\_data\\_up\\_to\\_date](https://www.wikidata.org/wiki/Wikidata:Data_donation#4._Keep_the_data_up_to_date)

recorded on the **related changes** page on Wikidata.<sup>18</sup> Users can check the revision history of a certain item and reverse it to the previous version if the change made by others is wrong. The recent changes can also be tracked by Wikidata's API service.<sup>19</sup> Besides, individual users can track changes by using the **watchlist** function and can set up an RSS or Atom feed to receive updates about items on a watchlist.<sup>20</sup>

---

<sup>18</sup><https://www.wikidata.org/w/index.php?hidebots=1&hidecategorization=1&target=Property%3AP497&showlinkedto=1&namespace=0&limit=500&days=30&title=Special:RecentChangesLinked&urlversion=2>

<sup>19</sup> <https://www.mediawiki.org/wiki/API:RecentChanges>

<sup>20</sup> <https://www.wikidata.org/wiki/Help:Watchlist>

## [Workshop – 1]

### 歴史学におけるデータ共有，統合化，多角的協働

#### *Data Sharing, integration, and multi-proxy collaboration in historical studies*

人文学や社会科学におけるオープンアクセスとオープンデータの実現は、歴史資料の多角的・総合的なアプローチの促進につながります。日本の歴史研究は、多様なデータセットに対する大規模で長期的・学際的な種々の研究プロジェクトにより、過去 100 年にわたって発展し続けています。しかしながら、研究手法の標準化、メタデータの資源化や長期保存、災害時における資料データの活用などは、十分に実施できているとは言えません。研究資源として歴史資料や研究データを積極的に共有するための基盤についても、具体的にどのような整備が必要となるのか、社会的な課題に対応して整備する必要があります。本ワークショップでは、各種研究データへの十分なアクセスと長期保存・活用を実施するためにはどうすればよいのか、また、研究データリポジトリとしてどのような活用が見込まれ、利用を促進するためにはどうすればよいのかについて、各研究機関で実践されている事例をもとに検討します。ご関心がおありの方はぜひご参加ください。

A multi-faceted and integrated approach to historical resources leads to a synthetic discipline benefitting from open access and data in the humanities and social sciences. Japanese historical studies have advanced over the past century through large-scale, long-term, and multidisciplinary projects that integrate diverse datasets with sophisticated approaches. However, dispersed and heterogeneous data lead to technological challenges. Standardisation of methods, development of robust metadata, and long-term preservation and utilisation in case of disasters are insufficient. Sociological challenges, including inadequate rewards for sharing research and resource data, must also be resolved. This workshop examines how superior data access and preservation can be achieved to promote the attribution and acknowledgement of well-curated and federated data repositories. If you have any interests, please join in this workshop.

## 日程

ワークショップはJADHの第1日目の2021年9月6日(月) 10:30 - 14:00に開催します。

時間	タイトル	発表者
10:30-	開会挨拶 Opening	本郷恵子(東京大学史料編纂所所長) Keiko Hongo (Director, Historiographical Institute The University of Tokyo)
10:40-11:00	東京大学における日本史史料の長期利用とデータ共有・連結化 Long-term Utilization, Data Sharing, and Linking of Japanese Historical Materials by the University of Tokyo	渋谷綾子, 大向一輝, 山田太造, 中村覚, 渡邊要一郎, 平澤加奈子, 山田俊幸(東京大学) Ayako Shibutani, Ikki Ohmukai, Taizo Yamada, Satoru Nakamura, Yoichiro Watanabe, Kanako Hirasawa, Toshiyuki Yamada (The University of Tokyo)
11:00-11:20	全国遺跡報告総覧: 日本考古学の最大規模のデータベース SORAN: A Comprehensive Database on Japanese Archaeology	高田祐一, ヤナセ・ペーテル(奈良文化財研究所) Yuichi Takata, Peter Yanase (Nara National Research Institute for Cultural Properties)
11:20-11:40	歴史文化資料の保存・継承に向けたネットワーク構築とデータ連携の展望 Constructing international university network to preserve local historical resources	天野真志・後藤真(国立歴史民俗博物館) Masashi Amano, Makoto Goto (National Museum of Japanese History)
11:40-12:00	Trial for collecting metadata of research data in Asia アジアにおける学術データのメタデータ収集の試み	原正一郎(京都大学東南アジア地域研究研究所), 杉本重雄(筑波大学), 亀田堯宙(国立歴史民俗博物館) Shoichiro Hara (Center for Southeast Asian Studies, Kyoto University), Shigeo Sugimoto (University of Tsukuba), Akihiro Kameda (National Museum of Japanese History)
12:00-13:00	昼休憩 Break	
13:00-14:00	ディスカッション Discussion	ディスカッサント: 大向一輝, 山田太造 Discussants: Ikki Ohmukai and Taizo Yamada
14:00-	終了 Closing	

## [Workshop – 2]

## 海外 DH 教育動向調査

*Workshop on the state of Digital Humanities Pedagogy abroad*

昨今、人文学分野の学部・大学院教育における情報リテラシー涵養の必要性が叫ばれ、コースやカリキュラムの設置・導入が進んできています。このような事情に鑑み、2020年11月に発足した日本デジタル・ヒューマニティーズ学会「人文学のための情報リテラシー」研究会では、人文学およびデジタル・ヒューマニティーズ教育・研究に有用な情報リテラシー習得のためのカリキュラム編成を考案したいと考えています。本ワークショップでは、2000年代からこのような教育体制の整備を進めてきた欧米の事例を紹介いただき、国内の人文学教育・研究における情報リテラシーのあり方を展望する機会としたいと思います。ご関心がおありの方は、ぜひご参加ください。

## 日程

ワークショップはJADHの第1日目の2021年9月6日（月）15:00 - 17:00に開催します。

時間	タイトル	発表者
15:00-15:10	開会 Opening	小風尚樹（千葉大学人文社会科学系教育研究機構 助教） Naoki Kokaze (Chiba University)
15:10-15:30	提供者視点の取り込みのススメ Recommendation of taking provider perspective	福山樹里（国立国会図書館 利用者サービス部政治史料課占領期資料係） Julie Fukuyama (National Diet Library)
15:30-15:50	実践しながら考えるデジタル人文学：北米DHカリキュラム考察 Learning by Doing: North American DH Curricula	横山説子（シンガポール工科大学 デジタル人文学助教授） Setsuko Yokoyama (Singapore University of Design and Technology)
15:50-16:10	フランスにおける人文情報学教育の理念と実践 Principles and Practices of DH Education in France	長野社一（千葉大学人文社会科学系教育研究機構 特任研究員） Soichi Nagano (Chiba University)
16:20-17:00	議論 Discussion	司会：永崎研宣（人文情報学研究所首席研究員） Kiyonori Nagasaki (International Institute for Digital Humanities)