

Online Event

JADH 2023

Possibilities for Data-Driven Humanities

2023.9.20 Wed. – 22 Fri.



Organized by
Organizing Committee,
Japanese Association for Digital Humanities

Hosted by
National Institute of Japanese Literature

Co-organized by
International Institute for Digital Humanities



— International Symposium —

Exploring Possibilities for Data-Driven Research
in East Asian Studies

Admission
free!

Online Event

2023.9.21 Thu.

Hosted by
National Institute of Japanese Literature



Proceedings of JADH conference, vol. 2023

Edited by the JADH Program Committee and Local Organizing Committee

Copyright © 2023 by the Japanese Association for Digital Humanities

Published by the JADH Program Committee and Local Organizing Committee

20 Sep 2023

5-26-4-11F, Hongo, Bunkyo-ku, Tokyo, Japan

<https://www.jadh.org/>

Online edition: ISSN 2432-3144 Print edition: ISSN 2432-3187

Program Committee

Co-chairs

Natsuko Yoshiga (Osaka University, Japan)

Kiyonori Nagasaki (International Institute for Digital Humanities, Japan)

Paul Arthur (Edith Cowan University, Australia)

Marcus Bingenheimer (Temple University, USA)

James Cummings (Newcastle University, UK)

J. Stephen Downie (University of Illinois, USA)

Maciej Eder (Pedagogical University of Kraków, Poland)

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)

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 (National Institute of Japanese Literature, Japan)

Asanobu Kitamoto (ROIS-DS Center for Open Data in the Humanities / National Institute of Informatics, Japan)

Naoki Kokaze (Chiba University, Japan)

Chao-Lin Liu (National Chengchi University, Taiwan)

Yoko Mabuchi (Wayo Women's University, Japan)

Charles Muller (Musashino University, Japan)

Hajime Murai (Future University Hakodate, Japan)

Chifumi Nishioka (National Institute of Informatics, Japan)

Ikki Ohmukai (University of Tokyo, Japan)

Geoffrey Rockwell (University of Alberta, Canada)

Martina Scholger (University of Graz, Austria)

Masahiro Shimoda (Musashino University, Japan)

Raymond Siemens (University of Victoria, Canada)

Tomoji Tabata (Osaka University, Japan)

Ruck Thawonmas (Ritsumeikan University, Japan)

Toru Tomabechei (International Institute for Digital Humanities, Japan)

Kathryn Tomasek (Wheaton College, USA)

Ayaka Uesaka (Osaka Seikei University, Japan)

Raffaele Vighianti (University of Maryland, USA)

Christian Wittern (Kyoto University, Japan)

Taizo Yamada (University of Tokyo, Japan)

Hilofumi Yamamoto (Tokyo Institute of Technology, Japan)

Local Organizers

Keisuke Unno (National Institute of Japanese Literature, Japan) - Chair

Kazuaki Yamamoto (National Institute of Japanese Literature, Japan)

Shunsuke Kigoshi (National Institute of Japanese Literature, Japan)

Kuninori Matsuda (National Institute of Japanese Literature, Japan)

Nobuhiko Kikuchi (National Institute of Japanese Literature, Japan)

Noriko Matsubara (National Institute of Japanese Literature, Japan)

Taekjin Lee (National Institute of Japanese Literature, Japan)

Pre-event

Workshop: 「DHデータ基盤としてのデータセット～利用と提供から考える」 （第3回人間文化研究機構DH研究会）

DHを推進するためには基盤的なデータセットの存在が欠かせません。人間文化研究機構各機関のコーパス（国語研）、古典籍データ（国文研）、歴史地名辞書（機構本部）だけでなく、国内外の研究機関からもさまざまな基盤的なデータセットが公開されています。一方で、これらのデータセットの多くは、研究者または研究機関が中心となって構築されるケースが多く、これらを利用する立場の意見や他の研究者・研究機関が作るデータセットとの連携が十分には意識されてきませんでした。

本企画では、これらの基盤的なデータセットの利用者と作成者との対話を通じて、データ構築の在り方やデータセットの維持・提供および連携にかかる研究機関の役割などについて模索し、DHのより高度な基盤形成につなげる議論を行います。

日程	2023年9月20日（水）10:00～13:00
開催形態	Zoom
言語/Language	日本語
参加費	無料

プログラム

時間	話題	話題提供者
10:00-10:05	趣旨説明	関野 樹（人間文化研究機構DH推進室／国際日本文化研究センター）
10:05-11:25	データセットの利用事例の紹介と要望（20分×4件）	自身の研究やツール・データ構築での利用、利用者から見たデータセットの特色、他の研究資源やデータとの連携、提供者側への要望・質問など <登壇者> ・画像データ 鈴木親彦（群馬県立女子大学） ・言語資源データ 中俣尚己（大阪大学） ・時空間データ 北本朝展（国立情報学研究所） ・テキストデータ 石田友梨（岡山大学）
11:25-11:35	（休憩 10分）	
11:35-13:00	ディスカッション（応答30～40分＋ディスカッション50分～60分）	<司会> 関野樹、宮川創（人間文化研究機構DH推進室／国立国語研究所） <登壇者> ・利用側 北本朝展、石田友梨 ・提供側 ・画像データ 海野圭介（国文学研究資料館）

	<ul style="list-style-type: none"> ・言語資源データ 中川奈津子（国立国語研究所） ・テキストデータ 金甫榮（渋沢栄一記念財団） ・提供機関 大井将生（人間文化研究機構DH推進室／人間文化研究機構本部）
--	---

主催：大学共同利用機関法人 人間文化研究機構 人間文化研究創発センターDH推進室、JADH2023実行委員会

Pre-event

Workshop: 「研究者とライブラリアンとの対話：データ駆動型人文学の推進に向けたラウンドテーブル」 / "Dialogue between DH scholars and Librarians: A Roundtable on Promoting Data-Driven Humanities" (held in Japanese)

データ駆動型人文学、そしてその前提としてのDH（Digital Humanities）の浸透を目指すうえで、図書館等の文化機関と研究者との間の連携は欠かせない。その一方で、図書館と研究者との間にはDHに対する認識や温度差が大きい。加えて、日本の事情として図書館にサブジェクトライブラリアン等の研究者的な職位がほとんど設置されていないケースも多いことから、DH研究者と図書館との連携が活発化していない現状にある。本ラウンドテーブルでは、データ駆動型人文学およびDHの推進をテーマに、研究者と図書館員等の文化機関職員との間で対話を行うことで、これからの日本におけるデータ駆動型人文学の研究推進に向けた継続的な議論のための礎としたい。

In striving to promote data-driven humanities and, underlying that, the spread of Digital Humanities (DH), the collaboration between cultural institutions such as libraries and DH scholars is indispensable. On the other hand, there is a significant difference in the perception and response to DH between libraries and scholars. Moreover, given the circumstance in Japan, where research-oriented positions such as subject librarians are hardly established in university libraries, there is a current state where the collaboration between DH scholars and libraries is not being activated. In this roundtable, with the promotion of data-driven humanities and DH as the theme, we aim to establish the foundation for continuous discussions for advancing data-driven humanities in Japan in the future, by engaging in dialogue between DH scholars and cultural institution staff such as librarians.

日程	2023年9月20日（水）14:00～16:30
開催形態	Zoom
言語/Language	日本語
参加費	無料

プログラム

時間	話題	話題提供者
	司会・ファシリテータ	木越 俊介（国文学研究資料館教授）
14:00-14:10	趣旨説明・データ駆動型人文学研究に関する共有すべき前提	菊池 信彦（国文学研究資料館特任准教授）
14:10-14:25	DH研究者からライブラリアンへの期待と問題提起	永井 正勝（人間文化研究機構 人間文化研究創発センター 特任教授 国立民族学博物館／前 東京大学附属図書館アジア研究図書館U-PARL副部門長）
14:25-14:40	ライブラリアンからのリプライ	渡邊 由紀子（九州大学附属図書館／ライブラリーサイエンス専攻准教授）
14:40-14:55	DHとライブラリアンとの関係に関する英国の状況	堀野 和子（国文学研究資料館管理部 学術情報課 調査・管理係長 兼 データ標準化推進係長）／菊池 信彦
14:55-15:00	休憩	

15:00-16:30	参加者を交えたフリーディスカッション	
-------------	--------------------	--

主催: 大学共同利用機関法人人間文化研究機構 国文学研究資料館 古典籍データ駆動研究センター

Plenary and Keynote

JADH2023 Plenary and Keynote Session & NIJL International Symposium

"Exploring Possibilities for Data-Driven Research in East Asian Studies"

Date & Time	September 21, 2023 (THU) 13:30-15:00 (JST)
Venue	Online (Zoom)
Admission	Free of Charge
Hosted by	National Institute of Japanese Literature
Translatione	Japanese ⇄ English
Registration	<p>This keynote session is open to everyone and can be attended for free.</p> <p>Note that registration for the Keynote Session & International Symposium is separate from the main conference. Even if you are planning to attend the JADH2023, you will still need to separately register for this international symposium.</p>

Program

	<p>13:30-13:35: Opening Greetings Dr. Yasuaki Watanabe (Director of the National Institute of Japanese Literature)</p>
	<p>13:35-13:38: Explanation of Purpose Dr. Keisuke Unno (NIJL)</p>
	<p>13:38-13:58: Presentation 1 Problem-solving in Humanities through Data-Driven Approaches Dr. Keizo Oyama (NIJL)</p>

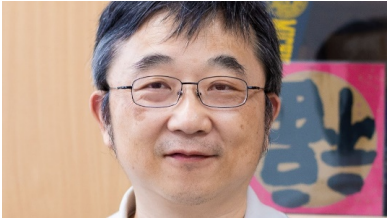



	<p>13:58-14:18: Presentation 2 Recent Development in Digital Humanities in Taiwan Dr. Chao-Lin Liu (National Chengchi University, Taiwan)</p>
	<p>14:18-14:38: Presentation 3 Teaching Data-Driven Humanities at Seoul National University and the University of Hong Kong Dr. Javier Cha (the University of Hong Kong, Hong Kong SAR, China)</p>
	<p>14:38-14:58: Q&A Session and Discussion - All participants</p>
	<p>14:58-15:00: Closing Remarks Dr. Keisuke Unno (NIIL)</p>

Photo by Felicia Buitenwerf on Unsplash

Hosted by National Institute of Japanese Literature

Long papers

Characterization by Dialogues in Hardy's novels: A Preliminary Quantitative Study of Three Works <i>Cao, Fanghui</i>	12
Developing an Automatic Classification Mechanism for Chinese Buddhist Texts Using Deep Learning Methods <i>Huang, Shu-Ling; Wang, Yu-Chun; Hung, Jen-Jou</i>	13
Can we conduct language documentation online? The development and application of Digital Platform for collecting Online language data (DOLD) <i>Lui, Pun Ho; Lai, Yik Po</i>	15

Short papers

Exploring the hidden Treasures of Geopark Heritage Resources through Data-driven Research <i>Aliakbari, Farzaneh; Tamborrino, Rosa</i>	19
Towards 'Linked Open SVOD Data': a data-driven study of the digital anime market <i>Delanaux, Remy; Roth, Martin</i>	19
Interactive Music Analysis Tool (I-MaT) <i>Eck, Sebastian Oliver</i>	20
Translations with cultural differences: A comparison study of original texts, and English translation texts of the Chinese novel using topic modelling <i>Fang, Wan-Zhen; Huang, Ling-Yi</i>	23
Applying text mining techniques to analyze the online news discourse on renting for elderly people <i>Hsiao, Yi-chen; Shao, Hsuan-lei</i>	24
A Quantitative Analysis of the Relationship between Physical Expression and Humor Creation in Rakugo <i>Kawase, Akihiro; Kinami, Chieri; Adachi, Junji</i>	27
Where did you come from, where did you go? Approaching cross-cultural heritage data for ancient evidence <i>Landau, Victoria Gioia Désirée</i>	28
Reconstructing and Interpreting the Historical Events of the White Terror Period with the Use of Generative Artificial Intelligence <i>Lin, Nung-yao; Hung, I-mei; Lin, Shu-Hui</i>	29
Using Results of Machine Learning as an Evidence for the Stylometric Analysis of Classical Chinese Poems <i>Liu, Chao-Lin; Mazanec, Thomas J.</i>	30
Database of Writing Systems and Orthographies for Okinawan Language: Toward Preservation of Okinawan Linguistic Cultural Heritage <i>Miyagawa, So; Carlino, Salvatore</i>	33
Interactive Storytelling with 3D Visualization for Illuminating the Impact of War in Ukraine <i>Morozov, Mykola; Kitamoto, Asanobu</i>	35
Digital Data Integration using Semantic Web and OPENAI <i>Moysaki, Georgia; Minadakis, Nikos</i>	38
Data Modeling and Visualization toward the Construction of 3D Platform for the Humanities <i>Ogawa, Jun; Ohmukai, Ikki; Nagasaki, Kiyonori; Kitamoto, Asanobu</i>	39
Affective Queer Narratives on Japanese Online Fora <i>Ohman, Emily</i>	42

Interactive presentations

Prototyping a Book Reading System with Overlaying Information Extracted by Large Language Models <i>Aubert-Bédouchaud, Julien, Maxime; Kitamoto, Asanobu</i>	45
A TEI-based Approach to Data Driven Analysis of Japanese Translationese <i>Camilleri, Gabriele</i>	47

Near-synonym noun-noun patterns in the Hachidaishu Dataset	
<i>Chen, Xudong; Hodošček, Bor; Yamamoto, Hilofumi</i>	49
Analysis of the Appearance Pattern Tendency of “Crying Scene” and Verification for Reproducibility of Categorization	
<i>Fukumoto, Takaki; Murai, Hajime</i>	52
Extracting “Darkness” in Contemporary Japanese Dark Fantasy	
<i>Kanazashi, Tomoya; Murai, Hajime</i>	54
DH Research Information Portal: A practice for “publicizing” DH methodology	
<i>Kikuchi, Nobuhiko</i>	55
Constructing fundamental behavior dataset for analysis and generation of story plots	
<i>Murai, Hajime; Ohta, Shoki; Ohba, Arisa; Fukumoto, Takaki; Aoyama, Mitsuki; Okuyama, Ryogo; Kanazashi, Tomoya; Saito, Yuni; Sato, Eiichi; Tomita, Masaki; Hodosawa, Tomowa</i>	57
Extracting the Relationship Between the Emotions Evoked in the Story and Acoustic Features of the Music	
<i>Okuyama, Ryogo; Murai, Hajime</i>	60
Online Reaction towards ChatGPT Ban from Education	
<i>Takagi, Miu Nicole; Ohman, Emily</i>	62
Development of a dataset for comparison between predicate verb phrases in the Kokinshu and their contemporary translations	
<i>Yamamoto, Hilofumi; Hodoscek, Bor; Chen, Xudong</i>	64

Panels

Digital Resources in Buddhist Studies in Taiwan – a Progress Report	
<i>Hsiang, Jieh; Hung, Jen-Jou; Hung, I-Mei; Ting, Pei-Feng; Lo, Hao-Cheng</i>	69
Possibilities of Digital Social Science and Data-Driven Studies	
<i>Shao, Hsuan-Lei; Huang, Sieh-Chien; Chao, Shiau-Fang; Yeh, Yu-Chun; Wu, Chia-Chia; Chang, Gia-Ming</i>	71
From Documents to DocuSky—Practice and Application	
<i>Tu, Hsieh-Chang; Hu, Chi-Jui; Kuo, Chih-Wen; Huang, Chia-Hung</i>	76

Long papers

Characterization by Dialogues in Hardy's novels: A Preliminary Quantitative Study of Three Works

Cao, Fanghui

u327503a@ecs.osaka-u.ac.jp
Osaka University, Japan

1 Introduction

Page (1973: 51) argues, "The dialogue in a novel is [...] multifunctional: it can serve to further plot, to develop character, to describe setting or atmosphere, to present a moral argument or a discussion on cabbages or kings, or to perform any combination of these purposes." However, the most important and the most productive one is the presentation and development of character.

Unlike the theater, the novel can't provide a observable personality of any character by the actor's performance (physique, costume, movements, facial and vocal qualities, etc), so these individual characteristics must be conveyed in words; and since dialogue is more dynamic compared with description or comment, it is often chosen to perform much of this task of characterization by novelists. Besides, when dealing with complex plots or extensive character-lists, the reader's memory needs artificial aids over a long period of reading experience, and that's the reason why the characterization by dialogues plays such an important role in novels.

Extracting the dialogue portion of a literary work can be a challenging task, particularly for novels with lengthy character lists and complex character relationships. Additionally, technical limitations have resulted in few previous studies examining the quantitative analysis of literary works' dialogue components.

Therefore, the goal of this study is to discuss how Hardy performed the characterization by the dialogues in his novels using quantitative analysis. This study focuses on the direct speech, which accounts for a significant proportion in the dialogues of Hardy's novels. The current paper considers whether there is a notable pattern could be recognized in the speech of his three representative novels, *The Return of the Native* (1878), *Tess of the D'Urbervilles* (1891), *Jude the Obscure* (1895), using the quantitative analysis approach that has been proposed in Cao (2023).

2 Methods

To prepare for the dialogue dataset, I collected the dialogues of each character in the three novels as follows. First of all, I selected the three books of Hardy in plain text from the Project Gutenberg, an online library of free eBooks. Secondly, I meticulously analyzed each utterance

in the novels through close reading and marked them up according to TEI (Text Encoding Initiative) guidelines. This resulted in three structured XML files, where I assigned attributes for the speaker and listener as "#who" and "#toWhom" within the <said> tags. Then, I parsed these XML files with lxml.etree in Python to extract the dialogue parts from the entire novel and represented the interaction network among the characters visually with networkx. After that, referring to the network, I analyzed the frequency patterns of the words used in the idiolects of the main characters with CasualConc to clarify the linguistic features and personality traits of them, as well as the relationships within the speakers. The table beneath shows the dataset collected in this study (Table 1).

Table 1. Basic data information of the three novels

* Tokens of Quotation Parts here means the tokens of contents enclosed by double quotations in the TXT files, which is the total of the tokens of contents enclosed by <q> tags and <said> tags in the XML files.

* Tokens of Direct Speech here means the tokens of contents enclosed by <said> tags in the XML files.

3 Results

After parsing the XML files, we have got three interaction networks of the novels (Figure 1, 2, 3). Thus, we can intuitively recognize the character relationships in the long-complicated stories. The characters in the centre of networks with larger nodes are the most important ones in the novel, while the characters that are pushed to the periphery are the ones with less importance.

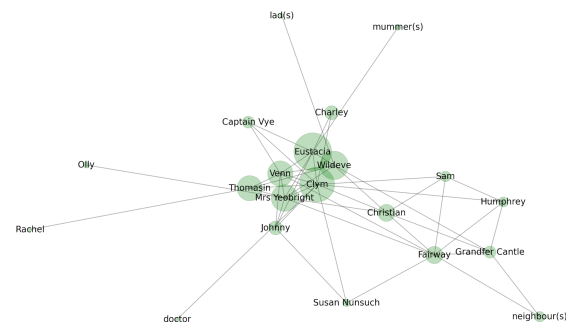


Figure 1. Visualization of the character interaction network in *The Return of the Native* (Characters must interact at least ten times to be included)

Figure 2. Visualization of the character interaction network in *Tess of the d'Urbervilles* (Characters must interact at least ten times to be included)

Figure 3. Visualization of the character interaction network in *Jude the Obscure* (Characters must interact at least ten times to be included)

Note that although most of the side characters have been removed from our network model successfully, there are still some collective nouns representing a group of supporting characters without specified names remained (*neighbour(s)*, *lad(s)*, *mummer(s)*, etc.). Hence, an additional condition, only those whose utterance is over 500 words are the research subjects, is added. Thus, we get a final character-list as presented in Table 2.

Table 2. Character lists of the three novels

As so far, we have got the final dataset ready and are allowed to go a step further by comparing the word usage of different characters. To examine the correspondence relationship between characters and words used in their utterances, this study utilizes correspondence analysis (CA). CA facilitates the visualization of analysis results through scatter plots, enabling an intuitive interpretation of the relationships between various categories present in rows and columns.

This study initially focuses on characters' usage of modal auxiliary verbs (*vm0). Figures 4 and 5 show the corresponding relationship between characters and modal auxiliary verbs. It was revealed that Hardy's rustics (right side) tend to use the abbreviation forms of modal verbs ("wo," "ll," "ca," "sha," "d") more frequently than his middle-class characters (left side), particularly those with better educational backgrounds on whom the dialectal influence is weakened by the influence of standard English. This finding aligns with Page's observations that "Hardy's dialogue [...] exhibits a wide variation of quality between the stilted and at times preposterous language of his middle-class characters, and the entirely different diction and rhythms of his rustics." (Page, 1973: 67).

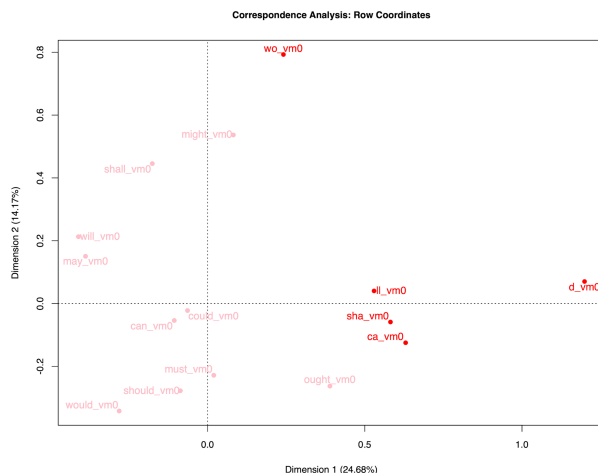


Figure 4. Correspondence analysis on modal auxiliary verbs in dialogues (By modal verbs)

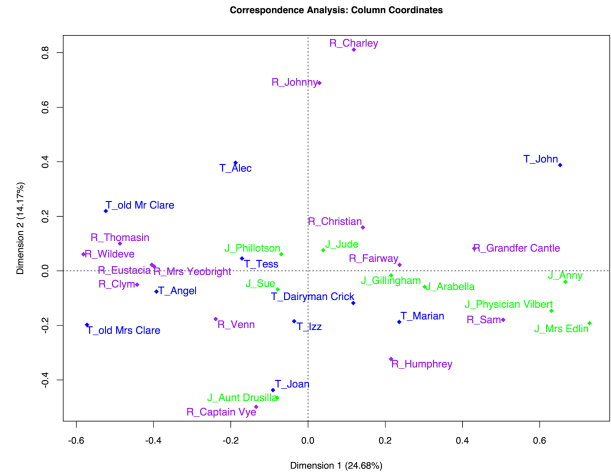


Figure 5. Correspondence analysis on modal auxiliary verbs in dialogues (By characters)

Bibliography

Cao, F. (2023). 『*Tess of the d'Urbervilles* の会話部によるキャラクターライゼーション』 *Tess of the d'Urbervilles no kaiwabu ni yoru kiyarakutaraizeishon* (Characterization by means of dialogues in *Tess of the d'Urbervilles*). Unpublished M.A. dissertation. Graduate School of Language and Culture, Osaka University.

Hardy, T. (1994). *Jude the Obscure*. <https://www.gutenberg.org/cache/epub/153/pg153-images.html> (accessed 13 March 2023).

Hardy, T. (1994). *Tess of the d'Urbervilles: A Pure Woman*. <https://www.gutenberg.org/cache/epub/110/pg110-images.html> (accessed 26 May 2022).

Hardy, T. (2006). *The Return of the Native*. <https://www.gutenberg.org/cache/epub/122/pg122-images.html> (accessed 13 March 2023).

Page, N. (1973). *Speech in the English novel*. London: Longman.

Developing an Automatic Classification Mechanism for Chinese Buddhist Texts Using Deep Learning Methods

Huang, Shu-Ling

d107104@dila.edu.tw

Dharma Drum Institute of Liberal Arts, Taiwan

Wang, Yu-Chun

ycwang@dila.edu.tw

Dharma Drum Institute of Liberal Arts, Taiwan

Hung, Jen-Jou

jenjou.hung@dila.edu.tw

Dharma Drum Institute of Liberal Arts, Taiwan

1. Introduction:

The Catalog of Buddhist Scriptures, known as Jing Lu (經錄), has played a crucial role in documenting and preserving the outlines of Buddhist texts in ancient China. With the advent of the Chinese Electronic Buddhist Tripitaka Collections (CBETA), a comprehensive database comprising over 2.3 billion words, the catalog has become an indispensable tool for research purposes. It not only facilitates access to the entire Buddhist canon but also establishes a standardized framework that enables comparative studies and research dialogues. However, the catalog faces certain challenges that need to be addressed. Firstly, there are contradictions within the traditional framework caused by the differences in genre-based or topic-based classifications. This inconsistency hinders the accurate categorization of texts and creates difficulties for researchers. Additionally, the continuous addition of new texts to the collection necessitates temporary categorization solutions to accommodate these additions. These challenges highlight the need for a more flexible and adaptable classification mechanism to ensure the effectiveness and relevance of the catalog in the face of evolving Buddhist literature. However, the traditional manual classification of Buddhist texts is time-consuming, laborious, and often lacks consensus. Thus this paper proposes the development of an automatic classification mechanism using deep learning methods.

2. Objective:

The primary objective of this study is to employ two deep learning methods to reclassify Buddhist texts in the last four relatively new categories of CBETA. These categories do not conform to the traditional thematic classification system and need to be integrated into the existing 19 categories. Moreover, the study aims to assess the suitability of the current text classification, address any ambiguities that may exist between categories, and discuss the reasons behind misclassifications. By conducting automatic classification on all texts in the first 19 categories, the research results will provide crucial reference data for manual revisions of the CBETA catalog.

3. Methodology:

In this study, three machine learning models are utilized: Bidirectional Long Short-Term Memory (BiLSTM) and BERT (Bidirectional Encoder Representations

from Transformers) serve as the primary models, while Support Vector Machine (SVM) is used as a baseline for comparison. BiLSTM is chosen for its ability to generate contextual word embedding and handle variable-length inputs, making it advantageous in text classification tasks. BERT, known for comprehending contextual nuances and dependencies, outperforms BiLSTM models in capturing semantic and syntactic relationships. As for SVM, its proven effectiveness and interpretability make it an ideal choice as a baseline model for comparison with the two deep learning models mentioned above. As expected, experimental results demonstrate BERT's superiority, achieving a test accuracy rate of 0.824, outperforming SVM and BiLSTM by 2.8 and 0.5 percentage points, respectively.

4. Findings:

The error analysis provides valuable insights into the classification process, revealing the following key findings:

(1) Higher effectiveness was observed in specific categories, including the Pure Land School Section (淨土宗部), Zen School Section (禪宗部), and Esoteric Section (密教部). Conversely, the Ratnakūṭa Section (寶積部) and Nirvāṇa Section (涅槃部) displayed lower effectiveness.

(2) The Suttanipāṭa Section (經集部) exhibited particular susceptibility to confusion due to unclear thematic boundaries and mixed categorization, emphasizing the need for adjustments to enhance classification accuracy.

(3) Certain categories tended to cluster together, such as the Nirvāṇa Section (涅槃部) and The Lotus Sutra Section (法華部); the Madhyamaka Section (中觀部) and Abhidharma-saṃnipāṭa Section (論集部); the Abhidharma Section (毘曇部) and Abhidharma-saṃnipāṭa Section (論集部); as well as the Zen School Section (禪宗部), Historical Biography Section (史傳部), and Cyclopaedia Section (事彙部). This clustering may be attributed to potential similarities in topics, a focus on human characters, or shared grammatical structures and linguistic styles. Further investigation is invaluable to explore these aspects in greater depth.

(4) The quantity of texts in different CBETA categories does not significantly impact the automatic classification performance.

(5) A recommended approach for reorganizing the final four new categories in CBETA is to combine the predictions from the three methods, considering the top-1 to top-3 results, which can then serve as a reference for human evaluation. For instance, considering the newly added book titled "Biography of Master Tsongkhapa" (宗喀巴大師傳 CBETA 2023.Q1, B11, no. 74). Properly categorizing this book falls under the No.18 Historical Biography Section. However, due to Tsongkhapa's association with esoteric teachings, BERT erroneously assigns it to the No.10 Esoteric Section, while BiLSTM correctly assigns

it to the No.18. This exemplifies the value of utilizing multiple predictions from different models to ensure precise categorization. If we consider the top three answers together, the accuracy rate can be increased to 90.3%.

(6) It is worth highlighting the potential consideration of cross-tagging for books that belong to multiple categories. This approach allows for a more comprehensive representation of the content and facilitates efficient retrieval and exploration of texts with overlapping themes or diverse subject matter. By exploring cross-tagging, the CBETA catalog can better accommodate complex texts and cater to the diverse needs of researchers and scholars.

5. Contributions:

This paper makes several significant contributions to the field. Firstly, it demonstrates the effectiveness of the deep learning method in classifying Buddhist Scriptures, surpassing traditional manual classification approaches. Secondly, it proposes the integration of the newer CBETA categories into the existing 19 categories, enhancing the organization and accessibility of the catalog. Thirdly, by analyzing the causes of misclassification based on the contents of Buddhist Scriptures, it explores the underlying connections between categories and provides guidance for future adjustments. Lastly, the findings of this study have been adopted by the Chinese Buddhist Electronic Text Association as a point of reference for manual classification.

6. Conclusion:

The utilization of deep learning methods, particularly BERT, in the automatic classification of Chinese Buddhist texts offers significant improvements in efficiency and accuracy. The research findings contribute to the refinement and enhancement of the CBETA catalog, bridging the gap between traditional practices and modern needs. The proposed automatic classification mechanism serves as a valuable tool for researchers, enabling them to navigate and explore the vast collection of Buddhist texts. Future research can further explore and expand upon the findings to advance the field of Chinese Buddhist studies and the digital humanities.

Key Words: CBETA, BERT, Dazhongjing, The Catalog of Buddhist Scriptures, text classification.

Bibliography

Devlin, J., Chang, M.W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805. <https://arxiv.org/abs/1810.04805>.

Li, F. H. and He, M. (2003). Research on the Chinese Tripitaka. Beijing Religious and Cultural Publishing Company.

Liu, G., and Guo, J. (2019). Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neuro computing* 2019, 337: 325-338.

Lu, C. (1980). Newly Compiled Catalog of the Chinese Buddhist Canon. Shandong Qilu Book Company.

Naseem, U., Razzak, I., Khan, S. K., and Prasad, M. (2021). A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5): 1-35.

Nguyen, B. and Ji, S. X. (2022). Fine-Tuning Pretrained Language Models With Label Attention for Biomedical Text Classification. arXiv: 2108.11809 [cs.CL]. <https://doi.org/10.48550/arXiv.2108.11809>.

Shi Huimin (2002). Introduction to the CBETA Electronic Buddhist Canon Integrated Catalog (CBETA Edition Catalog). *Information Management for Buddhist Libraries*, 32: 18-25.

Can we conduct language documentation online? The development and application of Digital Platform for collecting Online language data (DOLD)

Lui, Pun Ho

luiph001@gmail.com

The Education University of Hong Kong, Hong Kong S.A.R. (China)

Lai, Yik Po

laiyikpo@gmail.com

The Education University of Hong Kong, Hong Kong S.A.R. (China)

This study aims to explore a method to conduct remote language documentation. In the light of potential pandemic, the time and financial costs of field work, a remote approach to language documentation is required. To this end, this paper will introduce a web-based application for conducting language experiments and surveys online –

Digital Platform for Collecting Online Language Data (DOLD), which was newly developed by the Centre for Research on Linguistics and Language Studies, The Education University of Hong Kong, with the help of the JavaScript framework jsPsych (de Leeuw 2015). With DOLD, researchers can conveniently collect data from speakers of different languages around the world.

DOLD is an online platform in which the researcher can design their experiment(s), collecting linguistic data. Each experiment consists of task(s). Each task consists of stimulus and responses. Not only the actual responses are recorded, the response time is also filed.

Compatible with various types of stimulus materials (text, image, audio, video, and HTML content) and response data (text, audio, screen button, and keyboard, with reaction time), with a simple and flexible structure for researchers to design their own data-collection task package, DOLD can meet the research needs of different linguistic subfields, especially phonetics, clinical linguistics, language acquisition, psycholinguistics, dialectology, typology and language documentation.

An application of DOLD is collecting word list which is a common practice in in-person field work. First, researchers can upload photo and video stimuli onto the “media” section. Then, the researchers can create tasks that collect word list from languages. The task shows the stimuli (i.e., photo and video) and collect audio responses. Researchers can allow the participants to play back their recordings. If the participants are not satisfied with the recordings, they can re-record the audio, hence producing a more accurate pronunciation. By contrast, the playback function can be disallowed, aiming for the most natural recording (i.e., the first one) because the subsequent trials may be affected by participants’ awareness of producing “correct” expressions. It is optimal to design several tasks that aim at eliciting different tokens. E.g., One task for noun while another for verbs.

After creating the relevant tasks, researchers can create a new experiment by combining different tasks. The tasks can be freely added, removed and arranged. To distribute the experiment, send the generated link to the participants. The link can be dedicated to a specific participant, or can be available to public. After the experiment, in addition to the audio response, the response time from the participants is recorded. This may help assess whether the words are commonly used, that is, the duration of thinking to name the stimuli.

After a task package is set up in DOLD, no or minimal manual intervention from the researcher is required during the process of data collection. Participants can go through the process by themselves, at any time without making an appointment. Setting up an experiment or survey in DOLD and letting DOLD run it with participants without much manual intervention can be not only a convenient approach, but also a beneficial practice, which helps standardize the experiment or survey process. Sharing an experiment set up in DOLD with other researchers also allows easy and exact replication.

In terms of efficiency, DOLD reduces financial and time costs, compared to traditional field work. Traditional field work requires a researcher going to the locales, which are

sometimes rural areas, in order to collect language data. The money costs include the commuting cost (i.e., flight or car), accommodation, necessity and equipment. The time costs include the commuting time and the time of forging relationship with the target community. With DOLD, besides the cost of equipment, other costs are minimized to zero. On top of that, researcher’s risks and difficulties of going to different places are minimized.

In terms of application, DOLD can help collect linguistic data for cross-linguistically typological work (i.e., work on grammar). The traditional practices of conducting cross-linguistic work are by reading related grammar, and by consulting with the field workers of a given language(s). These methods may be biased by the field workers’ personal experience and intuition. To avoid the potential risk of inaccurate analysis, DOLD enables researchers to collect first-hand data by asking different language speakers to work on the same experiment. Same experiment is expected to yield a less biased work. Moreover, the source of data is more persuasive than personal communication.

Compared to other online experiment platforms that can be used for language documentation, DOLD is dedicated to language research, and is more transparent to use. One example is jsPsych, an open-source JavaScript library for developing web-based behavioral experiments (de Leeuw, 2015). While jsPsych helps researchers develop highly customized designs for the experiments, its utilization requires a working knowledge of JavaScript coding. Furthermore, jsPsych solely deals with the frontend aspects of an online experiment; additional knowledge is needed to manage backend components like server hosting and data storage. By contrast, DOLD does not require any knowledge of JavaScript coding, server hosting or data storage.

Another online tool for language documentation is the Gorilla Experiment Builder. It is a paid option that provides a graphical user interface for designing experiments, as well as backend infrastructure for data collection and storage. Designed to cater to a wide range of intricate behavioral experiments, Gorilla offers extensive features and details that go beyond the typical requirements of language experiments and surveys. It, as a result, appears heavy and complex to linguistic researchers, despite not requiring technical IT knowledge. On the contrary, DOLD is a free platform dedicated to linguistic and language researches. Aiming to provide a toolkit that focuses on the requirements of online language experiments and surveys, DOLD emphasizes simplicity and accessibility while offering necessary functionality.

Bibliography

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, 47(1), 1-12.

Short papers

Exploring the hidden Treasures of Geopark Heritage Resources through Data-driven Research

Aliakbari, Farzaneh

Farzaneh.aliakbari@polito.it
Politecnico di Torino, Italy

Tamborrino, Rosa

rosa.tamborrino@polito.it
Politecnico di Torino, Italy

The establishment of UNESCO Global Geoparks (UGGs) in 2015 was crucial to fulfilling the Sustainable Development Goals (Voudouris *et al.*, 2022). According to UNESCO definition 1, a Geopark is a single and unified geographical area with a geological heritage that are managed with a holistic concept of protection, education and sustainable development. Cultural heritage encompasses a broad range of tangible and intangible cultural elements, while Cultural Natural Heritage (CNH) represents the interaction between human communities and their natural environment, highlighting the relationship, beliefs, and practices associated with specific landscapes (Barrientos *et al.*, 2021; Tamborrino *et al.*, 2022). However, there is a need to broaden studies on geopark areas even if there are studies on protected heritage landscapes, the approach to composite geopark areas is not yet so clear where Cultural Natural Heritage (CNH) can constitute a lever for sustainable development. Therefore, the exploration of a range of data and information related to CNH characterization is needed to build a new approach. The current study identifies a non-European case study namely Qeshm Island UGG, and employs data driven approaches to investigate the history and CNH values. Qeshm Island is the Persian Gulf's and the Middle East's biggest island that is situated in the Strait of Hormuz in Iran (Sumanapala and Wolf, 2020; Nateghi and Bayat, 2021). Within this framework, the current study focuses on how to use digital tools to build experiences. Hence, the research gathers and analyzes the data about the CNH assets including geological data, historical records, cultural practices and traditions through surveys, and mapping processes using Geographic Information System (GIS). Finally, some insights regarding the application of digital tools to use the collected data for better understanding of the history and improving the visitor experiences are highlighted. The observations found in this study show that digital tools leverage collected data to provide a deeper understanding of history and enhance

visitor experiences through interactive and immersive approaches.

Notes

1. United Nations Educational, Scientific and Cultural Organization, <https://en.unesco.org/global-geoparks>, 2021

Towards 'Linked Open SVOD Data': a data-driven study of the digital anime market

Delanaux, Remy

delanaux.remy@gmail.com
Ritsumeikan University

Roth, Martin

roth1003@fc.ritsumei.ac.jp
Ritsumeikan University; Stuttgart Media University

While consuming digital video on the Internet was limited to free video sharing websites (such as YouTube or Dailymotion) in the 2000s, two separate branches converged into creating web-based video on-demand (VOD): native media creators wanting to profit off their creations, as well as traditional media companies from television and cinema wanting a new broadcasting channel for their productions and licenses. VOD on the Internet is an extension of pay-per-view (PPV) television and existing on-demand services shifting consumption from a fixed scheduling to an on-demand catalog, following what had already been happening with digital music (Lobato, 2018). In 2007, Netflix and Hulu merged subscription-based video on demand (SVOD) with digital platforms. For a monthly fee, unlimited access to a given media catalog is granted, instead of purchasing every desired title individually.

The proliferation of subscription-based models is not without issues for consumers, including the recent trend of newer platforms monopolizing content which could induce a necessity for an accumulation of subscriptions to access a majority of media content, in turn causing potential financial problems. Indeed, while some of these platforms' catalogs overlap, it also happens that some titles are not available anywhere; more specifically, they may not be available in a specific territory, or not available at that specific point in time (Lobato, 2018; Lotz *et al.*, 2022). License territoriality and temporality are growing concerns in a world where

digital media is heavily globalized, yet equal and stable access is not guaranteed. The lack of archiving solutions (whether in terms of the media itself, or the fact that it was available) creates concrete instances of inaccessibility that could lead to so-called "lost media" and to a "digital dark age", which represents a significant research issue (Kelly, 2022) and emphasizes the need for open solutions.

In this paper, we introduce our metadata-based approach to studying the spatiotemporal distribution of content across different SVOD platforms, organized in four steps: data acquisition, data cleaning, modeling, and analysis.

We argue that providing centralized, reliable, standardized resources for publishing data related to SVOD catalogs would be beneficial to both consumers and researchers. Any data model describing SVOD data needs to capture the spatiotemporal availability of media contents as much as is technically possible. We propose an abstract, simple data model and associated dataset to publish such data, using the Semantic Web and Linked Open Data frameworks and tools. Such a standardized, open model can be easily achieved and could reuse both existing models and existing data resources to be interlinked with. SVOD catalog data, despite being rarely studied, is relevant to study problems from fields ranging from Media Studies to Cultural Studies, as well as Platform Studies to observe the role of new media giants. This could also provide a good framework for working with transparent open data in Media Studies, thus perpetuating the hope for Open Science academic research and filling a notable gap in this field (Taurino, 2022).

In our study, we build and model datasets of SVOD catalogs for three major, global platforms (Netflix, Amazon Prime Video, and Disney+), allowing us to filter contents by their country of origin, as well as easily storing these catalog dumps in a simple format to see how these catalogs evolve in time and space. Using standards from the Semantic Web easily allow us to reuse existing vocabularies (or ontologies), compute statistics relevant to our problems of interest, and cross-query our computed data with existing media databases (e.g., the Media Arts Database from the Japanese government). To ensure interoperability with other databases, we reuse existing both general vocabularies (such as the W3C's OWL-Time ontology) as well as specialized vocabularies (such as the Dublin Core vocabulary or MovieLabs's ontologies for media creation and distribution).

As an example case study, this paper looks at the distribution of Japanese animated content (the core component of what is commonly called *anime*) in Japan and France. *Anime* content is discursively related to Japan (Suan, 2021), well-described in digital data resources (Pfeffer and Roth, 2020), globally conceived and distributed predominantly via digital channels such as VOD (Steinberg,

2012) and a very active media sector in those two countries (Petit et al., 2022). We analyze the availability of anime media on the major global SVOD platforms in France and Japan, and use our collected data to show how these contents are distributed in the two countries: availability over time, exclusivity vs. overlap, old vs. new, franchises vs. original creations, etc.

We argue that our approach offers a complementary perspective on contemporary media ecosystems, and thus on our globalizing media culture as a whole.

Bibliography

- Kelly, J. (2022).** "This Title Is No Longer Available": Preserving Television in the Streaming Age. *Television & New Media*, 23(1), 3–21.
- Lobato, R. (2018).** Rethinking International TV Flows Research in the Age of Netflix. *Television & New Media*, 19(3), 241–256.
- Lotz, A. D., Eklund, O. and Soroka, S. (2022).** Netflix, library analysis, and globalization: rethinking mass media flows, *Journal of Communication*, 72(4), 511–521.
- Petit, A., Steinberg, M., Crawford, C., Ristola, J., Altheman, E., Ciarna, S. and Berndt, V. (2022).** Anime Streaming Platform Wars: A Platform Lab Report. Concordia University Research Group.
- Pfeffer, M. and Roth, M. (2020).** Japanese Visual Media Graph: Providing researchers with data from enthusiast communities. *International Conference on Dublin Core and Metadata Applications*, 136–141.
- Steinberg, M. (2012).** *Anime's Media Mix: Franchising Toys and Characters in Japan* (NED-New edition). University of Minnesota Press.
- Suan, S. (2021).** *Anime's Identity: Performativity and Form beyond Japan*. University of Minnesota Press.
- Taurino, G. (2022).** Open-Data, Open-Source, Open-Knowledge: Towards Open-Access Research in Media Studies, *The Palgrave Handbook of Digital and Public Humanities*. Palgrave Macmillan, pp. 49–68.

Interactive Music Analysis Tool (I-MaT)

Eck, Sebastian Oliver

sebastian.eck@outlook.com

University of Music Franz Liszt Weimar, Germany

The Interactive Music Analysis Tool, I-MaT, (University of Music Franz Liszt Weimar, 2021a) is a modular program designed for producing visualizations, statistical analyses as well as tokenizations of textual music data. Its

functionalities are built on various commonly used python libraries, such as music21 (Cuthbert, 2023a) and MidiTok (Fradet *et al.*, 2021). I-MaT was developed as a sub-project within the fellowship project Computer-Assisted Music Analysis in Digital University Teaching (Pfleiderer, 2022; University of Music Franz Liszt Weimar, 2021c).

The fellowship project (2021) was located at the Institute for Musicology Weimar-Jena at the Franz Liszt University of Music Weimar, Germany.

Project Overview

The aim of the fellowship project “Computer-Assisted Music Analysis [...]” was to design, test, teach with and evaluate a comprehensive set of teaching modules for music analysis that make use of various computer-assisted, primarily quantitative analysis tools. The teaching modules were intended to complement conventional musicological and analysis courses and mainly focus on:

- “the computer-based annotation and visualization of sheet music texts and audio files,
- the statistical analysis of music corpora,
- the search for musical patterns (rhythms, melodies, harmony connections, etc.),
- and the comparison of interpretation” (Poliakov and Nadar, 2021a)

Within the project Computer-Assisted Music Analysis in Digital University Teaching, three computerized approaches to music score analysis were utilized, namely music21, CAMAT, and I-MaT (University of Music Franz Liszt Weimar, 2021d).

Music21 is a Python library created at the Massachusetts Institute of Technology (MIT), Cambridge for symbolic music representation and processing (Cuthbert, 2023b). CAMAT, on the other hand, is a Computer-Assisted Music Analysis Toolkit developed within the course of the fellowship project (Poliakov and Nadar, 2021b; Pfleiderer *et al.*, 2023). CAMAT uses its own unique data structure to overcome certain design problems found within the music21 framework (Poliakov and Nadar, 2021a). Tutorials for musicological university level courses were prepared and their usability evaluated; the tutorials include music examples and code that can be executed as Jupyter Notebooks (University of Music Franz Liszt Weimar, 2021b).

Objectives

As experienced during the fellowship project’s didactic test phase, working with Jupyter Notebooks and analyzing music with packages such as music21 or CAMAT proved challenging in particular for musicology students with none or only very limited knowledge of computer commands or programming languages.

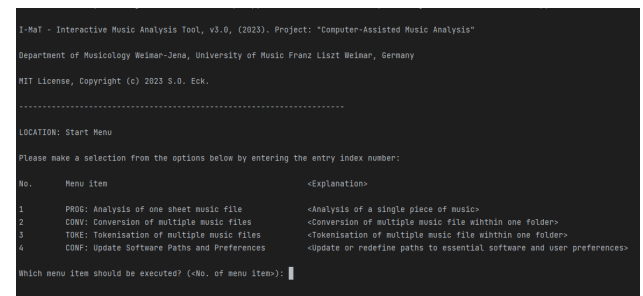
As an initial response, I-MaT emerged as a pragmatic solution, aiding students in quickly obtaining meaningful analysis results. Motivated by the success and the evident impact it had on the learning process, the development and refinement of I-MaT continued independently following the conclusion of the aforementioned fellowship project at the end of 2021.

Since then, this research endeavor has been guided by two primary objectives:

The first objective was to design an easily accessible tool with a well-structured interface, further lowering the barrier for musicology students to engage with methods from digital musicology.

The second objective was to further refine and modularize I-MaT’s program code, to create a modular and flexible tool that could effortlessly be extended by either adding new functionalities offered by already integrated modules (such as music21 and MidiTok), or by incorporating new python packages related to digital musicology or other relevant fields, e.g., computational linguistics.

I-MaT: Features and Accessibility



```
I-MaT - Interactive Music Analysis Tool, v3.0, (2023). Project: "Computer-Assisted Music Analysis"
Department of Musicology Weimar-Jena, University of Music Franz Liszt Weimar, Germany
MIT License, Copyright (c) 2023 S.O. Eck.

-----
LOCATION: Start Menu

Please make a selection from the options below by entering the entry index number:

No.      Menu Item                                     <Explanation>
-----
1        PROG: Analysis of one sheet music file         <Analysis of a single piece of music>
2        CONV: Conversion of multiple music files      <Conversion of multiple music file within one folder>
3        TONE: Tokenization of multiple music files     <Tokenization of multiple music file within one folder>
4        CONF: Update Software Paths and Preferences  <Update or redefine paths to essential software and user preferences>

Which menu item should be executed? (<No. of menu item>):
```

The Interactive Music Analysis Tool (I-MaT)

Addressing the steep learning curve often encountered with Python-based music analysis tools like music21 or CAMAT, the Interactive Music Analysis Tool (I-MaT) was designed specifically with user accessibility in mind. I-MaT utilizes a new and innovative approach to access, work with and implement various python libraries, such as, but not limited to, music21 or MidiTok for textual music analysis,

within one unified, user friendly text-based command-line-interface (CLI).

I-MaT allows users to quickly obtain results by navigating through simple dynamic menu structures and selecting methods and tools from predefined options (see Figure 1). The tool uses an accessible and easily extendable text-based CLI, with the underlying music21 and MidiTok commands remaining invisible to the user. While requiring minimal familiarity with command-line environments, the preference for a CLI over a GUI offers a compromise between user-friendliness and easy code extendibility.

This compromise was necessary to keep the barrier to entry as low as possible, and to encourage a broad usage and user-based participation in the tool's ongoing development via GitHub (Eck, 2023b). To further increase accessibility, I-MaT was distributed via Pypi.org (Eck, 2023c), allowing for an easy installation via integrated package-management systems such as the commonly used python pip installer. I-MaT's source code is complemented by an extensive online documentation that offers guidance to both users as well as contributors (Eck, 2023a).

While virtually encompassing all the functionalities of the integrated python packages for music information retrieval/tokenization, i.e., music21/MidiTok, I-MaTs functionalities are currently limited to a representative, yet well-tested set of statistical analysis, export, visualization, transformation and musical data tokenization tools.

With all those benefits at hand, I-MaT is a very flexible and powerful tool that could cater to the needs of a diverse range of users, from novice music analysts to advanced researchers.

Education and Training

In addition to its analytical capabilities, I-MaT serves as an effective didactic tool, further bridging the gap between musicology and the broader field of computer-assisted analysis and Music Information Retrieval (MIR).

I-MaT's various functionalities could provide valuable support for music and musicology courses at both high school and university levels. Its user-friendly interface and simple installation make it an ideal tool for students to quickly obtain results and explore various analytical approaches using computational methods, allowing them to gain a deeper understanding of the musical works they are studying.

Furthermore, I-MaT's modular design opens possibilities for launching educational projects centered around musicological programming, with the added advantage of seamlessly integrating their outcomes and functions into I-MaT through collaborative platforms like GitHub.

By using I-MaT, students can develop valuable analytical skills that are useful not only in musicology but

also in other areas of the humanities where data-driven methods are becoming increasingly important.

The Interactive Music Analysis Tool, I-MaT, should be seen as a contribution to Computational Musicology or Digital Musicology within the Digital Humanities.

Bibliography

Cuthbert, M. S. A. (2023a). music21. A Toolkit for Computer-Aided Musicology. <http://web.mit.edu/music21/> (accessed 19 July 2023).

Cuthbert, M. S. A. (2023b). music21. GitHub Repository. <https://github.com/cuthbertLab/music21> (accessed 19 July 2023).

Eck, S. O. (2023a). Interactive Music Analysis Tool (I-MaT). Dokumentation. <https://i-mat.readthedocs.io/en/latest/> (accessed 19 July 2023).

Eck, S. O. (2023b). Interactive Music Analysis Tool (I-MaT). GitHub Repository. <https://github.com/sebastian-eck/I-MaT> (accessed 19 July 2023).

Eck, S. O. (2023c). Interactive Music Analysis Tool (I-MaT). Pypi.org Distribution. <https://pypi.org/project/imat/> (accessed 19 July 2023).

Fradet, N., Briot, J.-P., Chhel, F., Seghrouchni, A. E. F. and Gutowski, N. (2021). MidiTok. A Python Package for MIDI File Tokenization. <https://archives.ismir.net/ismir2021/latebreaking/000005.pdf> (accessed 19 July 2023).

Pfleiderer, M. (2022). Computergestützte Musikanalyse. Fellowship für Innovation in der digitalen Hochschullehre. Abschlussbericht. https://stifterverband.org/file/10986/download?token=tl9S2EL_.

Pfleiderer, M., Poliakov, E. and Nadar, C. R. (2023). Analyze! Development and Integration of Software-Based Tools for Musicology Music Theory, Proceedings Innovation in Music 2022 Conference.

Poliakov, E. and Nadar, C. R. (2021a). CAMAT. Computer Assisted Music Analysis Toolkit. https://analyse.hfm-weimar.de/lib/exe/fetch.php?media=en:dmrn-16-proceedings_camat.pdf (accessed 19 July 2023).

Poliakov, E. and Nadar, C. R. (2021b). CAMAT. GitHub Repository. <https://github.com/Christon-Ragavan/CAMAT> (accessed 19 July 2023).

University of Music Franz Liszt Weimar (2021a). Fellowship Project Computer-Assisted Music Analysis. Interactive Music Analysis Tool (I-MaT). https://analyse.hfm-weimar.de/doku.php?id=en:interaktive_musikanalyse (accessed 19 July 2023).

University of Music Franz Liszt Weimar (2021b). Fellowship Project Computer-Assisted Music Analysis. Modules and Tutorials. <https://analyse.hfm-weimar.de/doku.php?id=en:tutorials>.

University of Music Franz Liszt Weimar (2021c). Fellowship Project Computer-Assisted Music Analysis. Project-Wiki. <https://analyse.hfm-weimar.de/doku.php?id=en:start> (accessed 19 July 2023).

University of Music Franz Liszt Weimar (2021d). Fellowship Project Computer-Assisted Music Analysis. Sheet Music Analysis. <https://analyse.hfm-weimar.de/doku.php?id=en:noten> (accessed 19 July 2023).

Translations with cultural differences: A comparison study of original texts, and English translation texts of the Chinese novel using topic modelling

Fang, Wan-Zhen

clairewzfang@gmail.com
Nanfeng College of Sun Yat-Sen University

Huang, Ling-Yi

lingyi0713@gmail.com
Mid Sweden University, Sweden

Introduction

Fortress Besieged written by Qian Zhongshu in the 1940's is a novel which describes the situation of Chinese intellectuals during the vicious Chinese Civil War. In 1961, a Chinese literature historian Hsia Chih-ting highly praised the novel's comic exuberance and acclaimed it as "the most delightful and carefully wrought novel in modern Chinese literature; it is perhaps also its greatest novel" (Hsia, 1961). With the English translation finished by Jeanne Kelly and Nathan Mao at the end of the 20th century, the author and his only novel have been enshrined as a classic in modern Chinese literature. Meanwhile, these studies on copious used metaphors or allusions manifest the long-term challenges in Comparative Literature (CL), which is characterized by dealing with the crossing borders in language. Both questions of "What is it?" and "How is it?" remain unclear and anxious in this discipline (Dominguez and Wang, 2016). Taking cultural differences into account, can we achieve Bermann's sense of "connection, relation, and dialogue" by translation in meeting the author's rich knowledge and rhetorical strategy?

Perspectives of Comparative Literature

Emily Apter invokes Derrida's notion of untranslatability and Reinhard's theory of "traumatic proximity" to abolish the divides of inside/outside, guest/host, owner/tenant in monolingualism (Apter, 2006). Apter

clarifies what can't be ignored is "systems theory formed part of a general post-war trend of "grand theory in the human sciences"" (Apter, 2009). Benefiting from the approach of Apter indicates the paper combines Latent Dirichlet Allocation with translated literary works to overcome the multiplicity of spaces in CL by manifesting some fundamental principles. Confronting the problems of language untranslatability, cultural differentiae, and the subjectivity of translator, we employed Lefevere's theory and Venuti's theory of foreignization to explain the transformation of cultural concepts. As E. A. Nida said, "The role of language within a culture and the influence of the culture on the meanings of words and idioms are so pervasive that cannot be disregarded" (Nida, 2001). Especially, Qian Zhongshu, the modern China's foremost "scholar novelist", entails an enormous challenge to the translators. To sum up, in line with Lefevere and Venuti's perspectives, we propose to employ topic models to reveal the differences and the varieties of the transmission.

Research method

Latent Dirichlet allocation (LDA) is a popular topic modelling technique to extract topics from a given corpus. LDA is a generative probabilistic model of a corpus. The basic idea is that texts are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words (Blei, Ng, and Jordan, 2003). In this study we built topic modelling of Chinese original texts and English translation texts respectively and we compared these two results according to (1) the labels of selected different numbers of the topics and (2) the labels of the best numbers of the topics.

Research results

We found that for Chinese original texts, the best numbers of the topics are two while for English translation texts, the best numbers of the topics are five (please see the link: <https://englishahq.neocities.org/FB1>). We matched the labels of English and Chinese novels. We also compared the coherence scores diagrams and found out that topics of English translations texts tend to be more dispersed than the topics of the Chinese texts. We found out that: (1) there are no matching topics between Chinese texts and English texts when the numbers of the topics are two and (2) there are no matching topics between Chinese texts and English texts when the numbers of the topics are five. Furthermore, we found out that the topics generated from the machine revealed that Chinese texts focus both on the descriptions of male and female protagonists while English texts focus more on the male protagonists. Lastly, for Chinese texts, the numbers of the topics tend to be more convergent when they are smaller while for English texts, the numbers of the topics tend to be more convergent when they are bigger. This could imply that the writing structures and the choices of the words were much more consistent and clearer in Chinese texts than those of the English texts.

Conclusions

We interpreted the research results mentioned above from two perspectives: First, we addressed Lefevere's theory of culture and translation to explain that English texts focus more on the male protagonists while Chinese texts focus both on the leading lady and man; second, we applied Venuti's theory of foreignization to illustrate the translators' adaptive choice-makings in English texts, which is more inconsistent and obscure than Chinese texts. In an introduction to English version, Nathan K. Mao, who is one of the translators, regards this novel as the worsening of the male protagonist's fortunes. Despite *Fortress Besieged* is known for the success of the courtship and marriage theme, Mao believes that the remarkable achievement of it is the shape of the main character, Fang Hung-chien. In other words, he thinks that Qian Zhongshu suggested the idea of besiegement by Fang and conveyed the condition of modern people. As Lefevere said that one of the professionals who controls the logic of culture is translator. The English texts topics focus more on the hero than on the heroine, which is the embodiment of the translator's ideology. Second, as scholars indicated that English translation texts are inclined to Chinese style. The translators chose word-for-word translation and disregarded the differences in culture and languages. They seized neither the subtlety in Chinese texts nor the conformity of English context, which led to the inconsistent and obscure of the English texts. Hopefully this study can contribute to appraising the value of translated version and exploring cultural differences by digital technology in the future.

Bibliography

- Apter, Emily.** (2006). *The Translation Zone: A New Comparative Literature*. Princeton: Princeton UP. pp.243-251.
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan** (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3: 993-1022.
- Dominguez, César and Ning, Wang** (2016). Comparative literature and translation: A cross-cultural and Interdisciplinary perspective. In Gambier, Yves and Doorslaer, Luc van. (eds), *Border Crossings: Translations Studies and Other Disciplines*. Amsterdam/Philadelphia: John Benjamins. pp.288-289.
- Hisa, C. T.** (1961). *A History of Modern Chinese Fiction*. New Haven, Conn: Yale UP. p.441.
- Lefevere, André.** (1992). *Translation, Rewriting and the Manipulation of Literary Fame*. London: Routledge. pp.13-16.

Nida, Eugene A. (2001). *Language, Culture, and Translating*. Shanghai: Shanghai Foreign Language Education Press. Preface.

Venuti, L. (2008). *The Translator's Invisibility: A History of Translation*. London: Routledge. p.24.

Applying text mining techniques to analyze the online news discourse on renting for elderly people

Hsiao, Yi-chen

yichen4682@gmail.com

National Taiwan Normal University, Taiwan

Shao, Hsuan-lei

hlshao@gapps.ntnu.edu.tw

National Taiwan Normal University, Taiwan

Abstract

The global trend of population aging and the rapid increase in the elderly population have made age discrimination in housing a significant issue in many countries. Taiwan is about to enter an ultra-aging society in 2025, facing not only a dramatic increase in long-term care needs and rising medical expenses but also a growing number of elderly individuals without self-owned housing who need to find suitable living arrangements. Therefore, the housing problem for the elderly is a major and crucial challenge for the Taiwanese government.

According to a report by Taiwan's TVBS news, less than 10% of landlords in the rental market are willing to provide opportunities for vulnerable tenants, and the number of landlords willing to accept elderly individuals living alone is even fewer (Chen Wen-Yue and Zhang Zhen-An, 2023). Elderly individuals face obstacles in renting housing, including poor living conditions and even homelessness. However, the current situation of the elderly rental housing problem is also reflected on social media. In the digital age, YouTube, as a video-sharing platform, differs from traditional one-way television broadcasting by providing interactive features and a high level of media richness. This study focuses on the public opinion on YouTube as its core, with research directions primarily concentrated on two dimensions of analysis: the content of news videos and platform comments. A total of 25 news videos were collected, spanning from March 2016 to April

2022, along with 1,189 comments, to explore the trends in public opinion regarding the issue of aging on social media platforms.

In terms of research methodology, we employed Python programming language for web scraping and utilized techniques such as sentiment analysis and word frequency analysis to uncover the underlying meaning in the text. The study found that on the YouTube platform, news reports depicted the elderly in a predominantly pitiful image. The majority of public opinion expressed positive and compassionate emotions, accounting for 42% of the overall sentiment. Negative emotions accounted for 25%, while neutral emotions constituted 31%. Through the text mining analysis conducted in this study, a more comprehensive description of the phenomenon surrounding the aging issue can be provided, promoting a deeper understanding of the public sentiment towards this topic among the general population.

Keywords: elderly housing rental, text mining, public opinion, image construction, framework theory.

Research Background

According to the latest statistical indicators from Taiwan's Ministry of the Interior, the aging index in Taiwan reached 147.06 in March of the 112th year (2023), which represents the ratio of the population aged 65 and above to the population aged 14 and below. The higher the index, the more severe the aging situation. The National Development Council estimates that Taiwan will enter a super-aged society by 2025. However, in an era of soaring housing prices, the issue of housing for the elderly has also come to the forefront. Despite government interventions such as promoting public housing and rental management, the core problem in the rental housing market cannot be directly resolved, and age discrimination in housing rentals for the elderly still persists. The Mama Foundation, which has been dedicated to promoting comprehensive reforms in Taiwan's rental housing market, recently collaborated with the Ministry of the Interior to propose measures for transparency in the rental black market. Through exchanges and discussions on relevant policies and systems, they aim to promote comprehensive management of rental properties, protect tenants' rights, and stabilize the housing situation for vulnerable tenants through the integration of resources by social welfare organizations. However, the successful implementation of these policies also relies on the rise of public awareness. Therefore, this study focuses on public opinion on the YouTube video platform and analyzes the trends in online discourse using text mining techniques. The research aims to describe the phenomenon of elderly housing rentals, capture the thoughts of the general public, and enhance the comprehensive understanding of the

government, relevant social welfare organizations, and the public regarding this issue.

The research questions are as follows:

What are the positive and negative sentiments expressed in public opinion in online news articles about the issue of elderly housing rentals? What is the frequency distribution of these sentiments?

What is the relationship between public opinion and media text on the issue of elderly housing rentals on the YouTube video platform? What is the distribution of word frequencies in these texts?

Research Method

This study utilizes text mining techniques, such as sentiment analysis and frequency analysis, to analyze the corpus.

Data Collection

First, considering the rich medium of audio and video provided by the YouTube platform, this study selects YouTube as the target platform. Before collecting the corpus, we conducted searches using keywords such as "elderly housing rentals" and "senior citizen rentals." We noticed a significant level of discussion in news reports related to this issue. Therefore, for this study, we focus on the comments left by YouTube viewers as the core of our research.

Through keyword searches, we found 25 news videos on the YouTube platform. The news reports cover the period from March 2016 to April 2022, resulting in a total of 1,189 comments. We used the Python programming language to write a web crawler. First, by importing the Selenium package, we simulated the scrolling behavior of users to capture the comments on the YouTube platform.

To further categorize and classify the news report videos, we analyzed the 25 selected videos on elderly housing rentals and assigned them categories based on their format and distinctive features.

Chinese Word Segmentation and Sentiment Analysis

In this study, we employ big data for sentiment analysis to avoid the errors caused by subjective human evaluations. We use Jieba, a Chinese word segmentation tool, to segment each comment into individual words. The segmented words are then cross-referenced with the NTUSD (National Taiwan University Sentiment Dictionary), which contains

positive, negative, and neutral terms, for sentiment analysis. We calculate the sentiment scores and determine the distribution of positive and negative sentiments for each video, allowing for further analysis of the frequency distribution of emotional terms.

Furthermore, the selected 25 videos primarily consist of online news reports from mainstream news channels such as CTV (China Television) and CTS (Chinese Television System). The purpose is to understand the image of the elderly constructed by mainstream media. Additionally, this study explores the correlation between the number of likes and the sentiment of comments, aiming to understand whether the audience holds a positive attitude towards the news reports and the resulting emotional sentiments.

Research content

This study divides into two main categories: the direction of public opinion and the framework of media text coverage.

The trend of public opinion

This study used to write crawlers, grabbed message content from 25 news videos, read positive and negative texts from the captured corpus, and obtained an emotional score through sentiment analysis, adding one to the positive vocabulary and subtracting one from the negative, and the results showed that there were 16 positive message sentiments greater than negative ones, 8 negative emotions greater than positive emotions, and one positive and negative report. From the perspective of positive emotions, the frequency of words such as gratitude, gratitude, and sadness is high, and the commenters mostly share the situation they have encountered as the largest case, accounting for 42% of the whole, however, in the negative emotions, it is from the perspective of abuse and criticism, words such as poor and obscene appear more frequently, negative emotions account for 25% of the whole, however, neutral emotions are 31%.

Media text reporting framework

Through random sampling and classification of news films according to the characteristics of the report, we found that the image of the elderly constructed by the report is mostly weak, and the news reporting framework is presented with social issues, in which the landlord is shaped as a negative image, such as "refusing to rent for fear of trouble" or "bad landlord renting out a poor elegant house", trying to construct the plight and misery image

of the elderly and strengthen the negative image of the landlord, but rarely reports from the perspective of the landlord. In addition, ninety percent of the news reporting forms are presented using soft thematic reports, with emotional appeals as the communication strategy, and only one proportion of news reports are hard instant reports.

Expected contribution

In this study, we analyze online public opinion from the perspective of YouTube message content, and we also classify the characteristics of news reports and videos to compare the core public opinion with the content characteristics of the film, so as to understand the public opinion, sentiment and word frequency distribution of the elderly issue. In addition, this study discusses the characteristics and framework of the reporting text, compares the correlation between the two, reflects the public opinion trend and emotional distribution of the issue, and the public emotional attitude generated under the reporting text, which can also show that the media constructs social reality, and this study can remind the media producers to abide by the principle of objective and neutral reporting in the construction of the issue, and to avoid aggravating the expansion of social problems with multiple reporting views. Finally, in Taiwan's elderly rental housing issue, with the efforts of various social welfare institutions and the government, there are policies implemented, this study provides a description of the public opinion phenomenon of elderly rental housing as a reference for relevant group.

Bibliography

- Chen Wen-Yue and Zhang Zhen-An.** (2023). *Elders living alone can't rent a house! 90% of landlords are reluctant to rent to the elderly.* TVBS news, <https://tw.stock.yahoo.com/news/%E7%8D%A8%E5%B1%85%E9%95%B7%E8%BC%A9%E7%A7%9F%E4%B8%8D%E5%88%B0%E6%88%BF-9%E6%88%90%E6%88%BF%E6%9D%B1%E4%B8%8D%E9%A1%98%E7%A7%9F%E7%B5%A6%E5%B9%B4%E9%95%B7%E8%80%85-145344905.html>
- Florian Primig, Hanna Dorottya Szabó and Pilar Lacasa.** (2023). *Remixing war : An analysis of the reimagination of the Russian-Ukraine war on TikTok.* *Frontiers in Political Science* .DOI 10.3389/fpos.2023.1085149
- Eun-Ju Lee, Yoon Jae Jang & Myojung Chung .** (2020) . *When and How User Comments Affect News .Readers ' Personal Opinion:*

Perceived Public Opinion and Perceived News Position as Mediators. Digital Journalism. <https://doi.org/10.1080/21670811.2020.1837638>.

A Quantitative Analysis of the Relationship between Physical Expression and Humor Creation in Rakugo

Kawase, Akihiro

kawase@dh.doshisha.ac.jp
Faculty of Culture and Information Science, Doshisha University, Japan

Kinami, Chieri

kinami.chieri@dh.doshisha.ac.jp
Faculty of Culture and Information Science, Doshisha University, Japan

Adachi, Junji

adachi.junji@dh.doshisha.ac.jp
Graduate School of Culture and Information Science, Doshisha University, Japan

Introduction

In large audience settings, speakers employ a wide range of physical expressions — including arm, leg, and body movements, as well as vocal intonations — to effectively communicate their messages. This is particularly true in the art of *Rakugo*, a traditional Japanese storytelling form wherein a single performer embodies multiple characters, skillfully utilizing physical expressions to narrate their tales in an engaging manner.

Nomura and Maruno (2006) conducted an analysis of variance on the performance devices of *Rakugo* performers to determine the loudness, speed, pitch, tickle emphasis, and facial expressions, reporting that humor may be created by the tendency to use each factor. However, how humor is created when performers play different characters has not been clarified. According to Wu (2005), people show an increased frequency of eye movement and little change in posture or body orientation during jokes. Sweetser and Stec (2016) found that comedians tend to indicate speaker alternation by shifting their gaze and head direction from

side to side when acting out characters. Furthermore, Logi and Zappavigna (2021) found rich physical expression to be associated with the creation of humor with respect to monomania in stand-up comedy. However, how physical expression leads to the creation of humor in situations such as *Rakugo*, in which one person plays multiple roles, remains to be clarified.

This study aims to elucidate the connection between physical expressions and humor generation in a single performer embodying multiple roles, and to identify the distinctions in physical expressions that humor creation hinges on, drawing on *Rakugo* performer's techniques from a quantitative perspective.

Analysis procedure

In this study, we manually extracted 90 seconds each of *Rakugo* performers acting out their characters from 14 videos included in the “Classical *Rakugo* Masterpiece Collection,” a historical video document. We annotated the videos using ELAN6.2 (Fig. 1) to extract the parts in which the physical expressions used by Logi and Zappavigna (2021) (e.g. Table 1) occur in the videos.

The number of occurrences and duration of each physical expression were measured based on whether or not the audience laughed. A residual analysis of chi-square test was conducted for the number of occurrences at a significance level of $\alpha=0.05$. A two-sample Wilcoxon test was conducted for the duration of occurrence. In addition, a logistic regression analysis was conducted to determine how each physical expression was related to laughter.

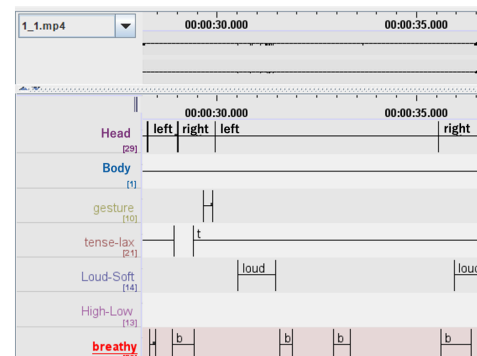


Fig.1: Example of the annotation work screen using ELAN6.2

Table 1: 20 physical expressions across 4 categories

Physical expression	Observed elements
---------------------	-------------------

Head orientation	The direction of the performer's head in relation to the audience: right; left; top; down; front.
Body orientation	Direction in which the performer's body is facing: right; left; up; down; front.
Gesture	Labels are assigned to symbolic gestures that do not directly express the content of speech
Voice quality	Tense; lax; rough; smooth; high; low; loud; soft; breathy. However, vibrato, plain, nasal, and non-nasal identified by Logi and Zappavigna (2021) did not occur in this study.

Results and discussion

The results of the residual analysis of the chi-square test revealed an association between the occurrence of laughter and physical expression. A two-sample Wilcoxon test was conducted on the total occurrence time of each physical expression, revealing no differences between the occurrence time and the absence of laughter for any of the physical expressions.

Table 2 depicts the results of a logistic regression analysis after selecting the explanatory variables with the lowest Akaike Information Criterion (AIC). The results confirm the tendency of *Rakugo* performers to create humor when their heads turn to the right. However, humor was less likely to be created when *Rakugo* performers used higher or softer voices, or when they turned their bodies to the right for longer periods of time. These physical expressions were related to the fact that *Rakugo* performers often tilt their heads to the right when they play the role of a boke (funny man). In addition, when a male character played a female character or intentionally used a whisper without emphasizing his words, humor was less likely to be generated due to the change in voice quality.

Table 2: Results of logistic regression analysis (after variable selection)

Number of occurrences			Duration time		
Physical expression	Estimate	P-value	Physical expression	Estimate	P-value
Head (right)	0.479	0.049	Body (right)	-19.399	0.048
Body (right)	-1.177	0.080	Body (down)	-11.383	0.057
Body (left)	1.203	0.071	Body (front)	-11.551	0.050

Voice (high)	-0.133	0.019	Gesture	1.888	0.067
Voice (soft)	-0.134	0.018			

Bibliography

Logi, L., and Zappavigna, M. (2021). Impersonated personae—paralanguage, dialogism and affiliation in stand-up comedy. *HUMOR*. <https://doi.org/10.1515/humor-2020-0023>.

Nomura, R., and Maruno, S. (2006). The Influence of Narrative Strategies Used by Rakugo Performers on Entertaining Effects. *Japanese Journal of Laughter and Humor Research*, 13, 13-23 (In Japanese). https://doi.org/10.18991/warai.13.0_13

Sweetser, E., and Stec, K. (2016). Maintaining multiple view- points with gaze. In B. Dancygier, W. Lu and A. Verhagen (eds.), *Viewpoint and the Fabric of Meaning*. Mouton de Gruyter, De Gruyter Mouton, 237–258.

Wu, Y. C. (2005). Frame-shifts in action: What spontaneous humor reveals about language comprehension. *Cognitive Science*, 17(2), 1–27, 2005.

Where did you come from, where did you go? Approaching cross-cultural heritage data for ancient evidence

Landau, Victoria Gioia Désirée

victoria.landau@unibas.ch

Digital Humanities Lab, University of Basel, Switzerland

Cultural heritage studies and projects have embraced aspects of digitization and computer-assisted approaches since decades, paving the way for «digital cultural heritage» and allowing for digital media, tools and infrastructures to support the preservation, documentation and dissemination of cultural (and natural) sites, objects and more intangible processes.

When connecting challenges faced by cultural heritage professionals in generating heritage data and translating it into a digital format for long-term use, with the disciplinary problems of those working with evidence from antiquity specifically, difficulties of the two both converge and diverge — issues of provenance, conservation and community meet concerns of periodization, contextualization and scholarship.

Two case studies illustrate these issues across millennia, evidence from Graeco-Roman Egypt and from Greco-Bactria. Alexander III of Macedonia (commonly known as Alexander the Great) defined the Eurasian continent in the 4th century BCE, venturing east, defeating the vast Persian Empire and conquering a territory roughly spanning from today's Greece to Pakistan. This immense expansion had longstanding consequences for the world he left behind after his death in 323 BCE.

In the south of his empire, the new Ptolemaic Kingdom would begin a period today broadly termed «Graeco-Roman Egypt», consolidating power by creating new deities combining Greek and Egyptian elements, and instating a dynastic cult deifying its rulers. Further, Greek became the administrative language and official *lingua franca*, causing the majority of the written evidence of late- and post-Pharaonic Egypt to be transmitted to modern times in Ancient Greek.

In the east, initially under the rule of the new Seleucid Kingdom, the region of Bactria seceded to become the Greco-Bactrian Kingdom. It experienced influences by the neighboring Maurya Empire, which over time promoted Buddhism in the region, and Bactria would subsequently expand into the «Indo-Greek Kingdom» following the fall of the Mauryas. This area of overlapping cultures, religions and territories gave rise to «Greco-Buddhist art», more specifically termed «Gandhara art», named after the valley and region where most of its development occurred.

While reconstructing the histories of ancient artefacts can prove difficult for archaeologists and ancient historians, following the path of ancient objects in more recent history has proven no easier. Egyptian objects, especially papyri (the main writing material of the time), from the Graeco-Roman period and earlier are spread across countless institutions, where they are often housed after having been purchased, e.g., during the «papyrus boom» at the turn of the 19th to the 20th century, or after being bequeathed following the passing of private owners. These private collectors with connections to institutions have most often been researchers themselves, such as Prof. em. Karl Kalbfleisch to the University of Giessen in 1953 or Prof. em. Hachishi Suzuki to Tokai University (Tokyo) in 2010.

Likewise, objects from Gandhara (modern-day Pakistan and Afghanistan) have made their way to collections far from their source, such as the Tokyo National Museum, the Ancient Orient Museum (Tokyo), the Metropolitan Museum of Art (New York), the Musée national des arts asiatiques Guimet (Paris) and the British Museum (London), and continue to be available on the antiquities market, sold at auction to private collectors.

Seeing the spread of these artefacts worldwide, the historical data extractable from surviving ancient evidence includes cross-cultural exchanges and syntheses in antiquity, as well as cross-cultural ownership and interpretation in

modernity. This presents researchers with the double-task of tracing the transfer of culture/cultural practice in antiquity, and also following the paths of objects/cultural artefacts in recent history. Due to disciplinary borders and traditions, these two aspects of research work are not always undertaken by the same professionals. Adopting a concept like «provenance», expanded from art history to broader historical, archaeological and archival fields of research, is an example for a successful integration into rather than an overhaul of established approaches and expertise, facilitating transdisciplinary implementation.

Different institutions, platforms and projects utilize very different approaches to this type of cross-cultural evidence in their care, and thus multiple questions arise about the proper handling of these artefacts and their data. Seeing as physical collections regularly face the dilemma of assigning objects to a predefined area in their inventory and exhibition spaces, can the digital sphere remedy the situation by e.g., placing these objects in multiple contexts at once? How should we go about constructing (meta)data for these cross-cultural artefacts, which may not neatly fit into pre-determined categories? How should collections, projects and platforms proceed to «label» their objects, with origins both ancient and modern? How can accessibility be guaranteed for researchers and the public alike, satisfying the needs of both? And how can we map and visualize the history of these objects compellingly, acknowledging failings of the past and the potential for the future (e.g., «digital repatriation»)?

This talk addresses some aspects common to these questions and discusses existing and emerging methodologies proposed to holistically meet the expectations that go hand in hand with analog and digital custodianship of (ancient) cultural heritage. These solutions can include establishing connecting points between individual collections, the development of umbrella resources designed to accommodate the content of different systems, and developing joint standards across fields — all to facilitate open, safe and long-term interactions between objects, data, institutions, projects, scholars and the public.

Reconstructing and Interpreting the Historical Events of the White Terror Period with the Use of Generative Artificial Intelligence

Lin, Nung-yao

nungyao@gmail.com

National Taiwan University, Taiwan

Hung, I-mei

yimay0519@gmail.com
National Taiwan University, Taiwan

Lín, Shu-Hui

siokhui@ntnu.edu.tw
National Taiwan Normal University, Taiwan

In Taiwan's White Terror period, human rights were greatly violated, and various sectors of society have been working hard to achieve transitional justice, commemorate victims, and rebuild social trust. To facilitate the transmission of these historical events and data to more people, many databases and platforms have been established, including the "National Human Rights Museum" 's "Historical Sites of Injustice Archive", the "White Terror Literature", "Human Rights Archives System" the "Archival Information System," and the "Taiwan Transitional Justice Database" built by the "Transitional Justice Commission." In this preliminary project, data sorting is through DocuSky's personalized service, focusing on the victims of the White Terror period, sorting out the related anthologies, and the memory of official files, private documents, and traces of the space field, reconstructing the social context of the white terror period, and a map of victims' life stories embedded with a microscopic perspective. With interpretation and recollection methods, it reproduces the social atmosphere of Taiwan during the period of Taiwan's White Terror, and the life stories of the victims and other related individuals.

In recent years, with the development of generative artificial intelligence (AI) technology, natural language processing (NLP) technology such as ChatGPT has been able to generate grammatically and semantically correct sentences and texts, as well as generate images and music. This study uses generative AI technology to reinterpret Taiwan's White Terror period through text, images, and music, based on the historical events that have already been sorted and stored in databases. The main purpose is to target children, who have fewer transitional justice resources, and provide them with a new interpretation and understanding of these events, which may seem distant and unfamiliar to them, through a visual-auditory-textual sensory approach. In this project, we take victims of White Terror, such as Mr. Cai Kunlin, as a case study. Using the memories of these parties, historical records, and literary and artistic depictions, we create a written narrative of the events of their suffering. With this narrative as the core content, we introduce AI audio-visual generation technology to reconstruct the historical context of the White Terror.

This paper aims to use various historical sources collected and organized by experts to reconstruct the

historical background and texts of the White Terror period. At the same time, this paper also applies AI technology to transform the sources into texts and graphics suitable for children's reading, in order to increase children's interest and understanding of history. The method of this paper is similar to creating a Doraemon's time machine, leading children to experience the historical scene of the White Terror firsthand. It helps them to establish their own value judgments and social responsibilities, thereby promoting the practice of transitional justice. This not only contributes to the understanding and memory of historical events by various sectors of society, but also helps to build a more just and humane society.

Using Results of Machine Learning as an Evidence for the Stylometric Analysis of Classical Chinese Poems

Liu, Chao-Lin

chaolin@g.nccu.edu.tw
National Chengchi University, Taiwan

Mazanec, Thomas J.

mazanec@ucsb.edu
University of California, Santa Barbara, USA

Background

When it comes to the stylometric analysis of literary texts, linguistics features are common and basic choices (Liu et al., 2018; Mikros, 2009). Given a collection of classical Chinese poems, we may calculate the frequencies of the bigrams in the poems, and show the observed statistics in Table 1. This table provides the 10 most frequent bigrams in poems of five Chinese dynasties in the collection. These five dynasties include Tang, Song, Yuan, Ming, and Qing. ¹ It is easy to perceive the impression that the most frequent bigrams in the dynasties are similar to each other. If we are further informed that these rankings are based a sufficiently large collection, i.e., that the collection contains more than 796 thousand poems and that there are at least 45 thousand poems for any of these five dynasties; then, one may be willingly to infer that the classical Chinese poems of different dynasties may share some common characteristics and that this is an interesting phenomenon that is worthy of further investigation.

Applying machine learning-based approaches offers an alternative perspective for researchers to find the common characteristics among literary collections. Consider the techniques of classification (Alpaydin, 2020). In a typical

two-category, say A and B, classification task, we provide sufficient training samples of categories A and B to a learning program, which aims to find ways to tell A and B apart based on the training samples. Then, we prepare another collection of objects of both categories A and B that are not seen by the learning program before, and we ask the leaning algorithm, which already learned, to determine the categories of the objects of the new collection. If the learned learning algorithm can achieve a high accuracy, then we may infer that the algorithm has learned ways to differentiate A and B based on some characteristics of A and B, which we probably might not be able to observe directly. Such a process indirectly shows that objects of category A have something in common, so do objects of category B.

A Classification Task of Classical Chinese Poems

We obtained 29558, 50974, 83442, 165717 poems of the Tang, the Northern Song, the Southern Song, and the Ming dynasties from Sou-Yun (n.d.) for the current experiments. ² Table 2 shows the numbers of poems of *pentametric quatrains*, *regulated pentametric octaves*, *heptametric quatrains*, and *regulated heptametric octaves* in these four collections. ³ We used a BERT model that was pre-trained with the texts of ancient Chinese (Wang and Ren, 2022) to convert the poems into dense vectors. The experience in the recent development of deep learning indicates that these BERT vectors are capable of capturing the unobservable semantic characteristics of texts, and using such BERT vectors led to excellent performances in tasks that requires the understanding of natural language texts.

We split these BERT vectors of these four types of poems in each of the dynasties into two subparts: 70% for training and 30% for testing within each subpart.

This time, we were doing six four-category classification experiments. We mixed the PQ, RPO, HQ, and RHO poems of two dynasties with labels PQ, RPO, HQ, and RHO, respectively. We had six experiments because there were six possible combinations when we had four dynasties in total.

Results of Empirical Evaluation

As we outlined above, we used the training part to train a classifier, and we chose to use the technique of *logistic regression* first. ⁴ Table 3 shows the results of the classification results. The results indicate that the classifier can differentiate these four types of poems, even though we have mixed these four types of poem from different dynasties together.

Tables 3 shows the confusion matrix for the experiment in which we mixed the poems of the Tang and the Northern Song dynasty. The F_1 measure for the PQ, RPO, HQ, and RHO were 0.85, 0.97, 0.96, and 0.97, respectively. Hence, the macro F_1 is their average, i.e. 0.94. The overall accuracy was 0.96. Table 4 summaries the performance measures of the six experiments.

Therefore, we may infer that the classifiers may have found some special characteristics about the four types of poems, and those characteristics remain rather stable from dynasty to dynasty. For otherwise, the classifiers should have performed poorly in the experiments.

Discussions

Stylometric analysis that bases on word-level features of the texts typically also relies on statistical analysis of the adopted features. An advantage of this statistical approach is that people might inspect and read the texts for verification, perhaps with the assistance of software tools. Some researchers have challenged such a procedure. Brennan and Greenstadt (2009) even attacked a neural-network based model that used linguistic features as its basis like (Matthews and Merriam, 1993). Our approach does not directly rely on linguistic features of the texts. Our current work was motivated by Underwood's recent work (Underwood, 2019), in which he had to work very hard to find features for his classifiers.

We actually built our classifiers with different classification techniques, including decision trees, random forest, gradient boosting, and support vector machines. Not all of these classifiers performed as well as the logistic regression models. Therefore, we did not have to worry very much about the possibility that the BERT vectors might encode the lengths of the poems, for otherwise the other types of classification techniques could have detected such hints as well.

That not all of the classifiers performed very well has an important logical implication: when a classifier performs very well for the tasks, there might be a special way to consider that the characteristics of a type of literary text remain stable.

Acknowledgments and Brief Responses to Reviewers' Comments

We are obliged to the reviewers who provided important comments for this abstract. There is no denial that there are a lot of details that we can add to the abstract to make it a complete technical report. Given the 1000 word limitation for the main text (not counting this response section), we have attempted to outline the background, the poems, the machine learning procedures, the experimental results, and a brief discussion. We understand that the sketchy contours of our work cannot satisfy the needs of full-fledged reviews, yet we hope that we can discuss these issues with the reviewers and participants during the conference.

This exploration started when Liu visited the University of California, Santa Barbara as a visiting scholar and Mazanec was Liu's host at the time of this writing. It was Mazanec's idea to follow Underwood's examples (2019) to examine whether there existed hidden common features in the classical Chinese poems that were authored by poets who lived across a number of dynasties. Liu conducted the experiments and wrote this extended abstract.

	1	2	3	4	5	6	7	8	9	10
Tang	何處	不知	萬里	千里	不可	今日	白雲	不見	春風	不得
Song	不知	春風	平生	人間	萬里	千里	不可	不見	何處	歸來
Yuan	萬里	春風	白雲	今日	不見	何處	青山	人間	風吹	千里
Ming	萬里	何處	千里	白雲	春風	青山	不見	不知	明月	不可
Qing	何處	萬里	不知	十年	千里	不見	春風	不可	風吹	東風

Table 1. Ten most frequent bigrams in the poems of five different dynasties.

	pentametric quatrains (PQ)	regulated pentametric octaves (RPO)	heptametric quatrains (HQ)	regulated heptametric octaves (RHO)	total
Tang	2338	12484	7267	7469	29558
Northern Song	3176	13710	16187	17901	50974
Southern Song	6013	19518	33172	24739	83442
Ming	12303	42795	43312	67307	165717

Table 2. The quantities of the four types of poems in the four dynasties.

Tang & Northern Song		predicted categories			
		PQ	RPO	HQ	RHO
true categories	PQ	2720	270	182	4
	RPO	155	13454	55	46
	HQ	348	161	15289	389
	RHO	5	251	208	17437

Table 3. The confusion matrix for the experiment in which we mixed the poems of the Tang and the Northern Song dynasties. How to read the table: The classifier classified 2720 PQ poems as PQ poems, and mis-classified 270, 182, and 4 PQ poems as ROP, HQ, and RHO poems, respectively.

	F ₁					accuracy
	PQ	RPO	HQ	RHO	macro	
Tang & Northern Song	0.85	0.97	0.96	0.97	0.94	0.96
Tang & Southern Song	0.85	0.96	0.96	0.97	0.94	0.96

Tang & Ming	0.90	0.98	0.97	0.99	0.96	0.97
Northern Song & Southern Song	0.88	0.98	0.97	0.98	0.95	0.97
Northern Song & Ming	0.92	0.98	0.97	0.99	0.97	0.98
Southern Song & Ming	0.93	0.98	0.98	0.99	0.97	0.98

Table 4. The performances of the six experiments in which poems of two dynasties were mixed at a time.

Bibliography

Alpaydin, H. (2020) *Introduction to Machine Learning*, fourth edition, The MIT Press, Cambridge, MA, USA.

Brennan, M. and Greenstadt, R. (2009) Practical attacks against authorship recognition techniques, *Proceedings of the Twenty-First Conference on Innovative Applications of Artificial Intelligence*, 60–65.

Liu, C.-L., Mazanec, T. J., and Tharsen J. R. (2018) Exploring Chinese poetry with digital assistance: Examples from linguistic, literary, and historical viewpoints, *Journal of Chinese Literature and Culture*, 5(2):276–321, The Duke University Press, USA.

Matthews, R. A. J. and Merriam, T. V. N. (1993) Neural computation in stylometry I: An application to the works of Shakespeare and Fletcher, *Literary and Linguistic Computing*, 8(4):203–209.

Mikros, G. K. (2009) Content words in authorship attribution: an evaluation of stylometric features in a literary corpus, *Studies in Quantitative Linguistics*, 5, 61–75.

Sou-Yun (搜韻): <https://sou-yun.cn/>

Underwood, T. (2019) *Distant Horizons: Digital Evidence and Literary Change*, Chapter 2, The University of Chicago Press, Chicago, USA and London, UK.

Wang, P. and Ren, Z. (2022) The uncertainty-based retrieval framework for ancient Chinese CWS and POS, *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, 164–168.

<https://huggingface.co/Jihuai/bert-ancient-chinese>

Notes

1. Tang : 唐, Song: 北宋 and 南宋, Yuan: 元, Ming: 明, and Qing: 清

2. We separated the Song dynasty into the Northern Song and the Southern Song dynasties, and did not consider the Yuan dynasty here.
3. pentametric quatrains: 五言絶句, regulated pentametric octaves: 五言律詩, heptametric quatrains: 七言絶句, and regulated heptametric octaves: 七言律詩
4. The logistic regression function was implemented by the scikit learn <<https://scikit-learn.org/>>.

Database of Writing Systems and Orthographies for Okinawan Language: Toward Preservation of Okinawan Linguistic Cultural Heritage

Miyagawa, So

so-miyagawa@ninjal.ac.jp

National Institute for Japanese Language and Linguistics (NINJAL), Japan

Carlino, Salvatore

nanajuu@gmail.com

Daito Bunka University, Japan

Introduction

The Okinawan language, a Northern Ryukyuan dialect belonging to the Japonic language family, traces its roots to Okinawa Island in southern Japan. The estimated number of speakers in 2011 was 95,000; however, this figure is declining. The language has a rich literary history, appearing in various forms, such as Ryukyuan poetry, inscriptions at religious sites, royal tombs, and Bettelheim's 1855 Bible translations. However, its use was banned in educational institutions and public spaces after the annexation of the Ryukyu archipelago into the Empire of Japan in 1879. This suppression continued after the Second World War and the return of the Ryukyu Islands to Japan in 1972.

Despite its challenging history, the Okinawan language is experiencing a revitalization, with advocates pushing for the inclusion of Okinawan language education. Various writing systems reflecting modern pronunciation have emerged, including Hiragana syllabary-only, Katakana syllabary-only, Chinese characters combined with Hiragana syllabary, and the Roman alphabet. The Database of

Okinawan Writing Systems (DOWS) project aims to utilize various materials, from classical spellings to recent textbooks and phrasebooks, to compile all existing Okinawan language writing systems.

A Database of Okinawan Writing Systems

The classical writings of Okinawa, which include *Omoro Soshi*, a compilation of Ryukyuan poems, *Kumi Udui*, a type of traditional theater play, and religious songs, offer valuable insights into the culture, customs, and beliefs of the Ryukyu Kingdom, which ruled Okinawa between the 15th to the 19th centuries. These works are characterized by their poetic language, vivid imagery, and figurative meanings and are regarded as important cultural treasures of Okinawa. These pieces of classical literature and Bettelheim's translations of the Bible used traditional kana syllabary spellings. This traditional orthography no longer reflects the pronunciation of contemporary Okinawan.

We collected and analyzed eleven Modern Okinawan textbooks and dictionaries. Our analysis of the various orthographies revealed that these sources differ in their approach to writing consonant cluster onsets, which do not exist in Modern Standard Japanese, and in their renditions of syllables. We registered our findings on DOWS. Table 1 presents the textbooks and dictionaries we used together with examples of characters that represent variation.

	/wa/	/ʔwa/	/ja/	/ʔja/
NINJAL (1963)	ワ	ウワ	ヤ	イヤ
Nakamatsu (1999)	ワ	ワ	ヤ	ヤ
Nishioka et al. (2006)	ワ	ッワ	ヤ	ッヤ
Uchima & Nohara (2006)	ワ	ッワ	ヤ	ッヤ
Fija (2015)	わ	うわ	や	いや
Ogawa et al. (2015): Shuri Dialect	わ	'わ	や	'や
Ogawa et al. (2015):	わ	'わ / うわ	や	'や / いや

Tsukun Dialect					
Hanazono et al. (2020)	ワ	'ワ	ヤ	'ヤ	
Miyara (2021)	わ	っわ	や	っや	
Carlino (2022)	わ	'わ	や	'や	
Shimakutuba Seishohō Kentō Iinkai (2022)	ワ	ッワ / ?ワ	ヤ	ッヤ / ?ヤ	

Table 1: Eleven different orthographies for Okinawan DOWS is a sub-database of NINJAL Digital Archive (NINDA), built on Omeka S, a cutting-edge web publishing platform for digital cultural heritage projects. The DOWS project leverages the capabilities of linked open data (LOD) and the Resource Description Framework (RDF). Omeka S allows for the seamless integration of data and resources, providing a robust and user-friendly content management system. The use of LOD principles and RDF ensures that the Okinawan language data is widely accessible to interested parties around the world. By outputting RDF according to LOD principles, the DOWS project enhances the discoverability, reusability, and interoperability of the Okinawan language data.



Fig.1: Visualization of the DOWS database on NINDA (Omeka S)

This comprehensive database has enabled researchers to normalize and modernize spellings to make them typable and consistent with contemporary usage. The normalized spellings and major spelling variants will be added to an online database of an ongoing digital lexicography project called the Open Multilingual Online Lexicon of Okinawan (OMOLO).

Conclusion

In conclusion, the Okinawan language, facing the challenge of a dwindling number of speakers and a complex history of suppression, is now witnessing a resurgence in interest and efforts to preserve its rich linguistic and cultural heritage. By leveraging the power of digital humanities and modern technology, the DOWS project not only serves as an accessible vital resource for Okinawan language education and revitalization but also offers inspiration and guidance for the preservation of other endangered languages in today's increasingly digital and interconnected world.

Bibliography

- Carlino, Salvatore, 2022, “‘Nichiryū Shogo Online Jisho’ No Shōkai [JPN Introduction to ‘Japano-Ryukyuan Online Dictionary’],” *Nihongo No Kenkyū* [JPN: Studies of Japanese Language], 18(3), 52–59
- Fija, Byron, 2015, *Kimochi ga Tsutawaru! Okinawago Real Phrase Book: Pirin Paran Uchināguchi* [JPN: A Book of Real Phrases in Okinawan to Convey Your Feelings: Pirin Paran Uchinaaguchi], Tokyo: Kenkyūsha.
- Hanazono, Satoru, Satoshi Nishioka, Jō Nakahara, and Tomomasa Kuniyoshi, 2020, *Shokyū Okinawago* [JPN: Introductory Okinawan], Tokyo: Kenkyūsha.
- Miyara, Shinshō, 2021, *Uchināguchi Katsuyō Jiten* [JPN: Okinawan Practical Dictionary], Tachikawa: NINJAL.
- Nakamatsu, Takeo, 1999, *Okinawaken no Kotoba* [JPN: Languages in Okinawa Prefecture], Naha: Okinawa Gengo Bunka Kenkyūjo [JPN: Institute of Okinawan Language and Culture].
- NINJAL, 1963, *Okinawago Jiten* [JPN: Okinawan Dictionary], Tokyo: Zaimusho Insatsukyoku.
- Nishioka, Satoshi, Jō Nakahara, Noriko Ikari, and Yumi Nakajima, 2006, *Okinawago no Nyūmon: Tanoshii Uchināguchi* [JPN: Introduction to Okinawan: Enjoyable Uchinaaguchi], 2nd ed., Tokyo: Hakusuisha.
- Ogawa, Shinji, Hiromi Shigeno, Yūto Niinaga, Satomi Matayoshi, Nana Tōyama, Thomas Pellard, et al., 2015, *Ryūkyū no Kotoba no Kakikata: Ryūkyū Shogo Tōitsuteki Hyōkihō* [JPN: Writing Ryukyuan Languages: A Unified Orthography of Ryukyuan Languages], Tokyo: Kuroshio Shuppan.
- Shimakutuba Seishohō Kentō Iinkai, 2022, *Okinawaken ni Okeru “Shimakutuba” no Hyōki ni tsuite* [JPN: On Orthography of “Shimakutuba” in Okinawa], Naha: Okinawaken Bunka Kankō Sports-bu [JPN: Department of Culture, Tourism, and Sports, Okinawa Prefecture].
- Uchima, Chokujin, and Mitsuyoshi Nohara, 2006, *Okinawago Jiten: Naha Hōgen wo Chūshin ni* [JPN:

Okinawan Dictionary: Centering on Naha Dialect], Tokyo: Kenkyūsha.

Interactive Storytelling with 3D Visualization for Illuminating the Impact of War in Ukraine

Morozov, Mykola

mykola.morozov@tum.de

Technical University of Munich, Germany; National Institute of Informatics

Kitamoto, Asanobu

kitamoto@nii.ac.jp

ROIS-DS Center for Open Data in the Humanities; National Institute of Informatics

Introduction

Armed conflicts have long been a part of human history, leaving a trail of destruction and devastation in their wake. The war in Ukraine, which began in 2014, and the full-scale invasion, which started in 2022, have been no exception. The conflict has resulted in widespread destruction of infrastructure and homes. Such devastation can have a profound impact on individuals and communities, and educating people about these events is an invaluable step in ensuring the calamity does not repeat.

In modern times, information in different forms keeps being distributed through channels such as news articles, social media, and educational resources. While these methods provide valuable insights, they may not always capture the full extent of the impact of events like war.

To better understand and illuminate the aforementioned consequences, interactive storytelling and visualization technology can be utilized. By creating an immersive experience, users can gain a deeper understanding of the impact of war on local areas and regions. Furthermore, this approach ties in with the broader field of cultural heritage visualization and reconstruction, which seeks to preserve and reconstruct cultural heritage sites that have been destroyed or damaged.

Methodology

The use of interactive storytelling to share stories of disasters and their consequences has been explored in previous studies. It has proven itself to be effective at engaging and immersing users in a shown story, creating a strong emotional impact and improving memorization. (Traum et al., 2015) explored the effectiveness of interactive storytelling in teaching people about the impact of the Holocaust on survivors through an interactive conversation, increasing users' empathy and understanding of the tragedy. Similarly, (Vincent et al., 2015) found that 3D visualizations were an effective way to communicate the impact of disasters on destroyed cultural heritage sites and they could be achieved through crowd-sourcing or generating data. Building on these techniques in the domain of storytelling with 3D visualization (Thöny et al., 2018), we propose an app that generates 3D infrastructure views with contextual information.

The app has an architecture that is extensible, versatile, scalable, and performant to broaden its applicability. (Fanini et al., 2019) demonstrates a modular cloud-based architecture of web services, enabling archaeologists to reconstruct 3D cultural heritage sites from geospatial databases and to visualize them interactively. The app was designed as a distributed system, built on PlanetoidGen (Levus et al., 2022). It improves upon the processing "worker" software agent server architecture (Doran et al., 2010) by using a distributed message queue, and implements Morsel-driven query execution (Leis et al., 2014). The app has a full software stack, consisting of the document database MongoDB for storing 3D models and satellite images and the relational database PostgreSQL for storing geographic locations of buildings, running on the Kubernetes distributed system infrastructure. We also used OpenStreetMap Overpass as the map data source. Finally, we designed the front-end app using Unity for visualization and the back-end API for Unity to communicate with the database via gRPC and REST protocols.

Results

System features

The primary goal of the app is to allow teachers and students to study the procedural 3D views and publicly available images of landscapes damaged by the war in Ukraine using a before-and-after overlap slider. The app has two modes: a standalone client that generates new 3D models dynamically as a user explores areas on the map, and a web widget client focusing on a specific set

of landscape segments with predefined viewpoints and descriptions.



A view of the Novotoshkivske village in the application

These visualizations are accompanied by article sections providing context and outlining the specific events related to the depicted area (see Figure 1). People can explore these areas from multiple perspectives with features such as zooming, panning, and rotating the view.

Additional emphasis and visibility options exist (see Figure 2) to display and highlight roads, buildings, and satellite images. The app allows deeper integration with the surrounding article telling the story, allowing people to interact with it by navigating to a specific location by clicking a link in the text.

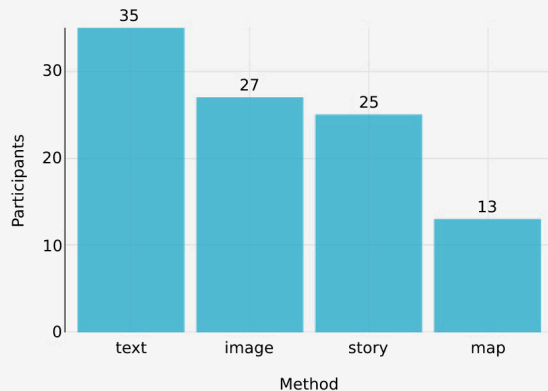


Different highlighting and visibility options for visualization of a large city in Eastern Ukraine, Mariupol

Evaluation

We measured the effectiveness and impact of information delivery of the app in comparison to other popular formats through a questionnaire of 25 participants. Although the number of participants is not large enough to reduce sampling bias, we tried to keep the diversity of participants by having Ukrainian, Russian, German and Japanese people, and having people with all genders and education levels between 14 and 33 years old.

PlanetoidGen QA Best



Questionnaire results showing the favorite news format based on participant feedback

Figure 3 shows the results of the questionnaire about their favorite news format among plain text, annotated images, labeled 2D maps, and the proposed interactive storytelling implementation. The app performed as expected, being chosen more often than maps and less than images (25% vs 13% vs 27% preference). The figure shows only the global data for the purpose of brevity, but trends were similar across different nationalities. In addition, Table 1 shows that the rating of the proposed system is on par with other forms of media. The result does not mean, however, that the app can replace all the benefits of traditional news formats; instead, we suggest that the app can be considered as an optional addition.

We also performed a test run in Lviv Physics and Mathematics Lyceum to obtain extended participant responses. We observed that the engagement and memory of students has improved after letting them control the app themselves. It suggests a correlation between interactivity and mental user involvement, with higher interactivity leading to an increase in enjoyment, engagement, and watch time. On the contrary, requiring the end user to perform manual analysis and complex understanding makes the experience less enjoyable.

Table 1. Feedback from the questionnaire about the news coverage app through interactive storytelling. The rating was given in the range [1-5] from worst to best

Method	Preference, %	Interest in topic	Overall rating	Preferred rating
Text	35%	3.54	3.34	3.40
Images	27%	3.44	3.89	4.04
2D Maps	13%	3.53	3.21	3.31

Storytelling	25%	3.72	3.99	4.08
--------------	-----	------	------	------

Conclusion

We developed an app for 3D visualization and interactive storytelling for the education of humanitarian aspects of the war. The modular design of the system allows the scientific community to adapt easily to other geospatial processing or visualization research. The evaluation shows that the app is effective in terms of emotional impact and engagement. The described features offer a more personalized and involved experience than traditional text articles, annotated images, and labeled 2D maps. By utilizing the proposed approaches, we hope to increase awareness of people on the subject of war in Ukraine and the impact of conflict on infrastructure and residents.

Bibliography

- Doran, J. and Parberry, I.** (2010). *Controlled Procedural Terrain Generation Using Software Agents*. IEEE Transactions on Computational Intelligence and AI in Games, 2(2): 111–19 doi:10.1109/TCIAIG.2010.2049020.
- Fanini, B., Pescarin, S. and Palombini, A.** (2019). *A cloud-based architecture for processing and dissemination of 3D landscapes online*. Digital Applications in Archaeology and Cultural Heritage, 14. Elsevier: e00100.
- Leis, V., Boncz, P., Kemper, A. and Neumann, T.** (2014). *Morsel-Driven Parallelism: A NUMA-Aware Query Evaluation Framework for the Many-Core Age*. Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. (SIGMOD '14). New York, NY, USA: Association for Computing Machinery, pp. 743–54 doi:10.1145/2588555.2610507. <https://doi.org/10.1145/2588555.2610507>.
- Levus, Y., Westermann, R., Morozov, M., Moravskiy, R. and Pustelnik, P.** (2022). *Using software agents in a distributed computing system for procedural planetoid terrain generation*. 2022 IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT) doi:10.1109/csit56902.2022.10000868.
- Thöny, M., Schnürer, R., Sieber, R., Hurni, L. and Pajarola, R.** (2018). *Storytelling in Interactive 3D Geographic Visualization Systems*. ISPRS International Journal of Geo-Information, 7: 123 doi:10.3390/ijgi7030123.
- Traum, D., Jones, A., Hays, K., Maio, H., Alexander, O., Artstein, R., Debevec, P., et al.** (2015). *New Dimensions in Testimony: Digitally preserving a Holocaust survivor's interactive storytelling*. Interactive Storytelling:

8th International Conference on Interactive Digital Storytelling, ICIDS 2015, Copenhagen, Denmark, November 30-December 4, 2015, Proceedings 8. Springer, pp. 269–81.

Vincent, M. L., Gutierrez, M. F., Coughenour, C., Manuel, V., Bendicho, L.-M., Remondino, F. and Fritsch, D. (2015). *Crowd-sourcing the 3D digital reconstructions of lost cultural heritage*. 2015 Digital Heritage, vol. 1. IEEE, pp. 171–72.

Digital Data Integration using Semantic Web and OPENAI

Moysaki, Georgia

georgia@advancesvs.com
Advance Services, Greece

Minadakis, Nikos

minadakis@advancesvs.com
Advance Services, Greece

Abstract: This document is a short description of the possibilities that are generated by Advance Services enterprise in the field of Digital Humanities and Culture, referring to written heritage sector. It includes a technical analysis of potential milestones of research e-infrastructures projects, whilst mentioned the implementation and results of PHAROS-Art Research Platform and Yashiro and Berenson Letters Platform.

Keywords: cutting-edge technologies, semantic web, OPENAI, Advance Services, integration, data, cultural heritage, complex querying answering

Digital Humanities Databases are scattered among various distributed heterogeneous infrastructures, keeping data and metadata with different formats prohibiting the integration of data on the same topic, period, person, artwork, and other entities of interest. This situation puts a variety of barriers in the conduction of scientific research. Scholars and researchers need days or even weeks in order to search, check and extract information from the relevant institutions for each case. The above mentioned phenomenon therefore generated the necessity of integration data solution. Efficient integration is necessary in every scientific domain. However, especially when the talk comes to humanities the integration becomes extremely meaningful, since data that has been gathered by different institutions, scientists in different places and on a timespan of hundred, if not thousands of years, must be collected and analyzed as a whole to allow complex scientific question answering.

To this end, we have designed and implemented a complete and comprehensive approach of methodology in order to harvest, normalize, clean, map, transform, ingest, integrate, reconcile, query and display humanities metadata and data using semantic models and repositories. This workflow is facilitated by an open source bigdata workflow execution and monitoring system (Apache NiFi:<https://nifi.apache.org/>) that allows the asynchronous, parallel, and efficient data management and manipulation.

On details, harvesting gathers periodic datasets in csv (https://en.wikipedia.org/wiki/Comma-separated_values), tsv (https://en.wikipedia.org/wiki/Tab-separated_values), or xml (<https://en.wikipedia.org/wiki/XML>) format using RESTful APIs (<https://aws.amazon.com/what-is/restful-api/>) or OAI-PMH (https://en.wikipedia.org/wiki/Open_Archives_Initiative_Protocol_for_Metadata_Harvesting) clients, storing these data to a distributed file system. These data are cleaned and normalized, semantically and syntactically by Open Refine (Open refine website:<https://openrefine.org/>) and Custom Scripts (<https://www.sentinel-hub.com/develop/custom-scripts/>) to be mapped to a semantic data model based on CIDOC-CRM (CIDOC CRM website: <https://www.cidoc-crm.org/>) and its family of models to be transformed in RDF (<https://www.w3.org/RDF/>) by a custom-made transformation engine algorithm. The RDF files are being imported in semantic databases (e.g. blaze graph) and reconciled to external sources to enrich their content. For example, personal information and data is being enriched using ULAN (Union List of Artist Name (ULAN) :<https://www.getty.edu/research/tools/vocabularies/ulan/>) Documentation: <https://www.getty.edu/research/tools/vocabularies/ulan/ULAN-Users-Manual.pdf>) and places information is being enriched by GeoNames (GeoName website:<https://www.geonames.org/>). Indexing takes place using SOLR (Solr website:<https://solr.apache.org/>) or Elastic search (Elastic Search website:<https://www.elastic.co/>) in order to increase the querying efficiency. On top of the semantic repositories, Rest APIs are built that are consumed by the GUI (https://en.wikipedia.org/wiki/Graphical_user_interface) which is using systems like Arches (Arches website: <https://www.archesproject.org/graphs/>) or CKAN (CKANwebsite:<https://ckan.org/>) for the visualization of data.

A novel structured search mechanism that enables gradually to build complex querying and answering is also supported. This process has been replicated on various international projects by Advance Deep Tech Services (AdvanceServiceswebsite:<https://www.advancesvs.com/>), such as PHAROS ArtResearch platform (PHAROS:TheInternationalConsortiumofPhotoArchives) and Yashiro & Berenson Letter Database (<http://itatti.harvard.edu/berenson-library/>)

collections/photograph-archives). PHAROS is an international consortium that has as a main objective to integrate PHOTO archival information coming from 9 research institutions namely Villa I Tatti-The Harvard University Center for Italian Renaissance Studies (<https://www.biblhertz.it/en/photographic-collection/>), Bibliotheca Hertziana Max-Planck-Institut für Kunstgeschichte (<https://www.biblhertz.it/en/photographic-collection/>), Deutsches Dokumentationszentrum für Kunstgeschichte – Bildarchiv Foto Marburg (<https://www.uni-marburg.de/de/fotomarburg>), Frick Art Reference Library (<https://www.frick.org/library/photoarchive>), Paul Mellon Centre for Studies in British Art (<https://www.paul-mellon-centre.ac.uk/>), Getty Research Institute (<https://www.getty.edu/research/>), National Gallery of Art (<https://www.nga.gov/>), Federico Zeri Foundation (<https://fondazionezeri.unibo.it/en/photo-archive/zeri-collection>), Kunsthistorisches Institut in Florenz (<https://www.khi.fi.it/en/photothek/index.php>). Nowadays, the PHAROS project includes data for more than 1,600,000 artworks, 2,700,000 photographs, 110,000 artists and 7,000 photographers, giving free access in a broad spectrum of fields and contributing to efficient research efforts. Yashiro & Berenson Letter Database is a project dedicated to influential art historians in the West and in Japan, respectively. Yashiro Letters Database is result of an international collaboration between two institutions closely linked with Berenson and Yashiro: Villa I Tatti - The Harvard University Center (<https://itatti.harvard.edu/>) for Italian Renaissance Studies, Florence (founded by Berenson) and Tokyo National Research Institution for Cultural Heritage (https://www.tobunken.go.jp/index_e.html). Through this project, it was achieved the digital exhibition of the letters, the annotation of persons and places with letter content, the support of multilingualism (English & Japanese), the link of letter content to people and places as well as other resources (i.e. WikiData entities) and the exploration of letters with Facets.

Apart from the classical keyword search, faceting, filtering, semantic search and advanced search, the novel structured search allows queries such as: “Give me negatives of photographs of artworks created by students of Leonardo Da Vinci and teachers of EL Greco that are kept in Louvre and Frick Museums and created by the technique of oil on wood”. Such queries that connect data coming from different databases and archives would otherwise require days of research from the art researchers. However, nowadays, they can be answered in a few seconds.

Moreover, we are also recommending the usage of Open AI (Open AI website:<https://openai.com/>) and specifically the libraries OpenAI and Pinecone (<https://docs.pinecone.io/docs/openai>) to overcome the learning curve issue of how to use the interface of structured search with the formulation of questions in natural language (either

written or oral). We have implemented transformation scripts from RDF to CSV format, compatible with OpenAI input format, and we are importing these CSV files to Pinecone. Finally, to enable running the querying functionality of OpenAI we construct the necessary configuration files in Pinecone and OpenAI. This innovative step combines and completes the research e-infrastructure in a meaningful way, expanding the research capabilities, whilst achieving a greater level of efficiency. Considering this and until now, we have surpassed the R&D phase of this last recommendation and are now using it on production projects.

Bibliography

Minadakis N., (2023), Digital Data Integration using Semantic Web and OPENAI

No one. Everything is included in the text linking with URL.

Data Modeling and Visualization toward the Construction of 3D Platform for the Humanities

Ogawa, Jun

htjk6513kbbk@gmail.com

ROIS-DS Center for Open Data in the Humanities, Japan

Ohmukai, Ikki

i2k@l.u-tokyo.ac.jp

The University of Tokyo

Nagasaki, Kiyonori

nagasaki@dhii.jp

International Institute for Digital Humanities

Kitamoto, Asanobu

kitamoto@nii.ac.jp

ROIS-DS Center for Open Data in the Humanities, Japan

Introduction

In the digital humanities, there has been much recent discussion regarding the application of 3D technologies. For example, 3D Scholarly Editions explores the annotation and

display of scholarly information around 3D objects in an analogy with text editions (Schreibman and Papadopoulos, 2019), while HBIM, which is a BIM for historical or heritage studies (Diara, 2022), and Extended Matrix have been proposed to accurately reconstruct and document 3D architecture using reliable archaeological data (Demetrescu, 2018). Additionally, SCOTCH Ontology was designed to document 3D reconstruction processes using a Linked Data approach (Vitale, 2017).

Building upon these cases, our study proposes an extended data model that properly represents the 3D scholarly information required for the humanities. Our model achieves full Linked Open Data (LOD) representation of various types of data relevant to 3D objects, including sources, interpretations, annotations, etc. The main theoretical contribution of our model is the introduction of an interpretation-centric approach to 3D scholarly representation. While some previous studies have highlighted the importance of scholarly interpretations in constructing 3D data, they have not represented them as data entities independent from 3D objects. We then consider each interpretation as an independent data item to achieve a more comprehensive representation of it.

In addition to proposing the data model, we also develop a visualization system for the scholarly representation of our data. Given the limited number of 3D scholarly representations based on Linked Data and the lack of precision in structuring information seen in prior examples, our system serves as an important pioneering example of such practice.

Interpretation-Centric Model for 3D Scholarly LOD

Our model comprises three distinct layers, as depicted in Fig. 1. Although each layer is separate, they are strongly interdependent and interconnected. By separating these layers, we gain the advantage of being able to collect, organize, and create data for each process independently.

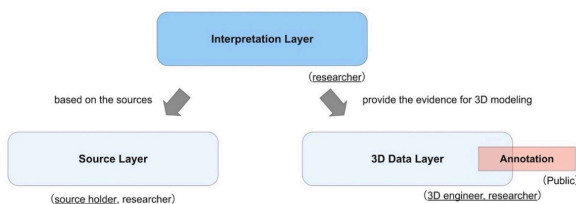


Fig. 1: Overall structure of the interpretation-centric model

The topmost layer, the Interpretation Layer, is the most important in the whole process of 3D scholarly representation as it holds the interpretative data used as parameters for constructing 3D models, including their shape, texture, material, date, archaeological features, etc. Thus, the fundamental concept of our model is that the 3D data should be constructed according to the interpretation, which is formed by referring to the sources. In this sense, the Interpretation Layer is really the axis of our data as it bridges the 3D Data and the related sources. Since interpretative parameters for 3D modeling can significantly differ according to the interpretation, they would be better separated from the 3D object data itself. Also, as this process would be mainly conducted by researchers, it would be preferable if the pro- and con- relationships between their interpretation can also be represented as data. Thus, we define the class `:Interpretation` shown in Fig. 2 to effectively implement such data organization.

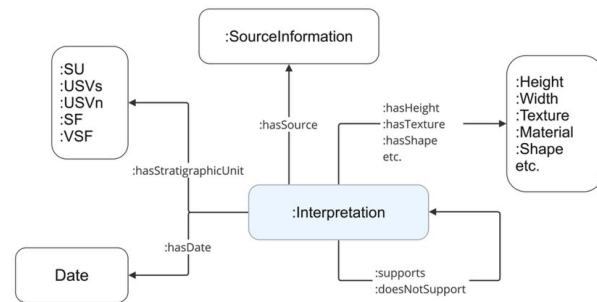


Fig. 2: Model for the Interpretation Layer

The Source Layer contains information on the sources used to form the interpretation. This layer aggregates the data related to a specific source, such as its title, author, source type, and web-linking information around an instance of the `:SourceInformation` class, as shown in Fig. 3. Collecting related sources would be carried out by source holders (individuals, museums, libraries, etc.) or researchers in the same way as ordinary archiving.

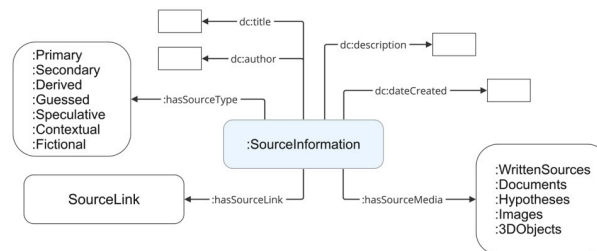


Fig. 3: Model for the Source Layer

We have defined the 3D Data Layer as the final layer in our model, which includes a 3D model along with its spatial context and various types of annotations as illustrated in Fig. 4. This layer comprises three key concepts: `:Object`, `:Space`, and `:ObjectGrp`. The `:Object` represents individual parts of a 3D space, such as columns or walls, while the `:ObjectGrp` includes all objects in the same model file. The `:Space` represents any semantically defined space made up of multiple objects, such as ‘temple’ or ‘corridor.’ Annotations can be added in different forms, including HTML pages, images, or TEI-encoded text data. The integration of TEI into the model is especially critical to represent text in a 3D context. Annotations are theoretically different from the interpretation because they do not necessarily function as evidence for 3D modeling process.

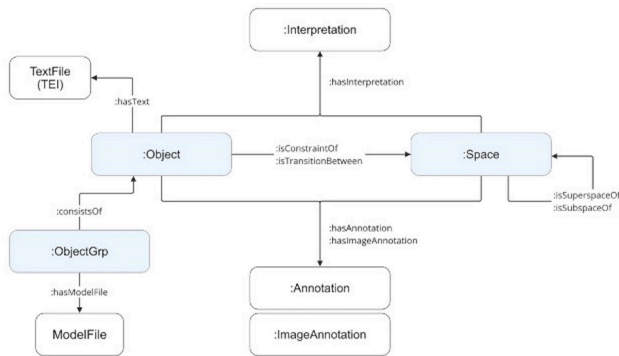


Fig. 4: Model for the 3D Data Layer

Pilot Platform for 3D Scholarly Representation

We are currently developing a pilot platform for searching, visualizing, and exploring 3D information using the data generated by the model described above. Since we have RDF data as the basis for the system, users can search for 3D models together with all the related information, which is described as Linked Open Data, with SPARQL queries and render them automatically to the 3D viewer shown in Fig. 5.

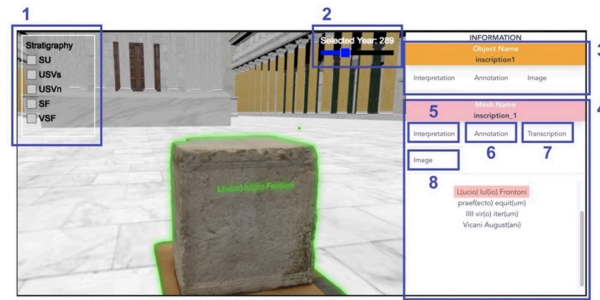


Fig. 5: The main view of the visualization platform

The 3D scene offers two interfaces, 1 and 2, allowing users to interactively visualize and filter 3D objects based on their date and archaeological categories, as described in the Interpretation Layer. The right-hand menu contains information panels for the Space (3) and Object (4) defined in the 3D Data Layer, each with multiple contents. The Object panel includes four content tabs: Interpretation (5), Annotation (6), Transcription (7), and Image (8).

Each content tab provides more detailed information about the object. In the Interpretation tab, users can view all the properties of the model described in the Interpretation Layer, as well as the sources in the Source Layer on which the interpretation is based. In the Transcription tab, textual information is displayed in both the panel and the 3D scene. As these transcriptions are automatically generated from TEI/XML data, all decisions made in the TEI encoding process can be reflected in 3D visualization.

Conclusion

The salient aspect of our research is the use of Linked Data to represent multi-layered information associated with 3D objects, particularly the incorporation of the Interpretation Layer. This approach enables us to efficiently consolidate, analyze, and visualize all related data, encompassing sources, interpretative data, annotations, and 3D text data, with a sound academic foundation, ensuring the quality of the discussion in a 3D context. In this regard, our research contributes to a global movement to establish a benchmark for the nascent field of 3D Humanities (Hendren, 2020).

Bibliography

Demetrescu, E. (2018) ‘Virtual Reconstruction as a Scientific Tool: The Extended Matrix and Source-Based Modelling Approach’, in Münster S., Friedrichs, K., Niebling, F. and Seidel-Grzesińska, A. (eds.) *Digital*

Research and Education in Architectural Heritage, Cham: Springer, pp. 102-116.

Diara, F. (2022) 'HBIM Open Source: A Review', *ISPRS Int. J. Geo-Inf*, 11(9), 472. doi: 10.3390/ijgi11090472.

Hendren, M. (2020) '3D Humanities: Digital Visualizations Promote New Research and Discourse', *National Endowment for the Humanities*, 8 May. Available at: <https://www.neh.gov/blog/3d-humanities-digital-visualizations-promote-new-research-and-discourse> (Accessed: 19 July 2023).

Schreibman, S. and Papadopoulos, C. (2019) 'Textuality in 3D: three-dimensional (re)constructions as digital scholarly editions', *International Journal of Digital Humanities*, 1, pp. 221-233.

Vitale, V. (2017) 'Rethinking 3D digital visualization: from photorealistic visual aid to multivocal environment to study and communicate cultural heritage', Ph.D. thesis, King's College London, London.

Affective Queer Narratives on Japanese Online Fora

Ohman, Emily

ohman@waseda.jp

Waseda University SILS, Japan

There is no doubt that Japanese society has become more accepting of sexual minorities and more supportive of their rights in the past few decades (Yamamura, 2023). Despite the increased acceptance only about 20% of sexual minorities in Japan are out to their friends and colleagues (Dehars & Iskandar, 2021) and only 1-8% are out to family out of fear of negative repercussions (Tamagawa, 2018). The fear of ostracism by family and friends after "coming out" and issues with self-acceptance as well as internalized LGBTphobia have been linked to higher suicide rates and *minority stress* both globally and in Japan (Saha et al., 2019; Komorida, 2021).

In this exploratory pilot study, we examine the attitudes present in online discussions related to sexual minorities. Specifically, we focus on the affective phrases used in connection with sexual minorities and compare this to a control group of assumed cis-gender heterosexual posts of an otherwise similar nature. We are using a dataset collected from online fora, mainly *Hatsugen Komachi*. We have to date recovered over 10,000 messages, both posts and comments, using search terms such as "LGBT" and "同性愛" (same sex love). Nearly all of them were in the "恋愛" (love/relationships) category and thus we collected a

random, assumed-cis-heterosexual, control sample from the same category.

We use a hybrid, iterative, approach where emotion lexicons are automatically adjusted for new domains with the help of affective word embeddings alongside large language models (Tohoku-NLP BERT) for contextual clues (Ohman & Rossi, 2023). We use emotion intensity rather than binary valence measures or emotion-association. Teodorescu and Mohammad (2023) showed that with the right bin size, lexicon-based methods produce results that are not only more human-interpretable, but also more accurate and real-world congruent than machine-learning-based models. Hence, we start with a lexicon, but enhance the lexicon by automatically adding words to the lexicon by using cosine similarity measures with specific proportional thresholds. We combine multiple Japanese emotion lexicons including JIWC and SNOW D-18 in the first step, but also use the WRIME (Kajiwaru et al., 2021) data together with BERT to separate the writers' and readers' emotions in the results (see Ramos et al., 2022).

The nature of the data, i.e., a communal help forum where people ask their peers for advice, means negative emotions are likely to dominate. Writers tend to intend to express more emotions (*anticipation* is the only exception), and readers seem to interpret the texts as containing relatively more sadness and anger, whereas disgust and other emotions are more evenly distributed between the readers and writers. The differences, however, for the most part are not statistically significant.

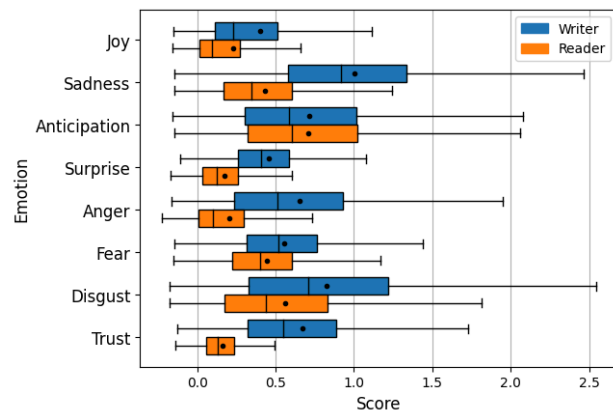


Fig. 1. Emotion scores for the sexual minority data.

It seems like there is little relational and relative difference between the two sets of reader and writer interpretations of the texts (fig. 1 & 2), but some in the intensity and the prevalence of affective content (fig. 3).

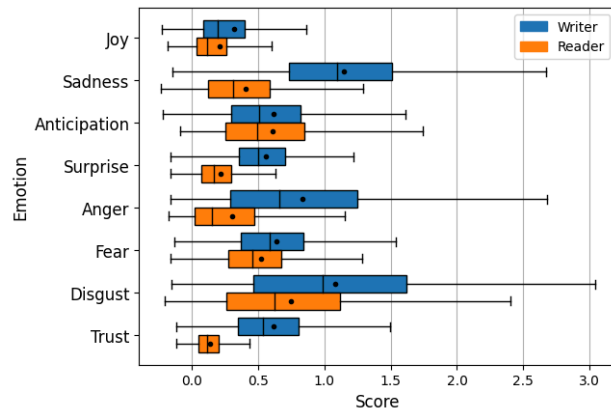


Fig. 2. Emotion scores for the control group (note that the x-axis goes to 3.0).

We also used simple emotion-word matching normalized by wordcount. Here the values are instances of emotion-associated expressions per 10,000 words where the intensity scores of values between 0 and 1 have been added together and the sum divided by the word count and then multiplied by 10,000 to generate comparable scores for both datasets. Figure 3 shows this normalized emotion word distribution in the data. Both the WRIME-based approach with BERT (fig. 1 & 2) and the enhanced lexicon-based approach (fig. 3) suggests the same thing: emotions are more overtly expressed in the control group. Log likelihoods indicate that the differences in the prevalence of affective language are statistically significant using this approach where emotion word distribution is the focus, however, unlike Saha et al. (2019) who found anxiety to be of high relative importance, we found the opposite and *anxiety* follows the same pattern as all other emotion-associations.

Fig. 3. Normalized emotion word distribution in the two datasets.

Overall, there are some minor differences between the two datasets that could possibly be linked to the theory of minority stress. The differences in affective expressiveness suggest that those writing about LGBTQ+ topics are choosing their words more carefully and avoiding overt expressions of emotion. The use of LGBTQ+ keywords to collect our data might also have shifted the valence of our data towards positive. We plan on collecting more data from a variety of sources using less explicit keywords to see if these findings can be made more robust, and to map affect over time to examine if there has been a change of attitudes in the past two decades that is reflected in the use of affective language in LGBTQ+ contexts.

Bibliography

Dehars, R.A.P., and Iskandar, K. 2021. "Company Policy VS Domestic: LGBT Discourse in Japan." *STRUKTURAL 2020: Proceedings of the 2nd International Seminar on Translation Studies, Applied Linguistics, Literature and Cultural Studies, STRUKTURAL 2020, 30 December 2020, Semarang, Indonesia*. European Alliance for Innovation, (2021).

Kajiwar, T. et al. 2021. *WRIME: A New Dataset for Emotional Intensity Estimation with Subjective and Objective Annotations*. In *Proceedings of the 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2021)*, pp.2095-2104, 2021.

Komorida, T. 2021. An Online Survey on the Mental Health of Lesbian and Bisexual Women in Japan. *The Senshu social well-being review*, 8, pp.65-78.

Ohman, E. & Rossi R. 2022. Computational Exploration of the Origin of Mood in Literary Texts. In *Proceedings of the 2nd NLP4DH workshop*. ACL Anthology.

Saha, K. et al. 2019. The language of LGBTQ+ minority stress experiences on social media. *Proceedings of the ACM on human-computer interaction*, 3(CSCW), pp.1-22.

Tamagawa, M. 2018. Coming Out to Parents in Japan: A Sociocultural Analysis of Lived Experiences. *Sexuality & Culture* 22, 497–520 (2018).

Yamamura, S. 2023. "Impact of Covid-19 pandemic on the transnationalization of LGBT* activism in Japan and beyond." *Global Networks* 23.1 (2023): 120-131.

Interactive presentations

Prototyping a Book Reading System with Overlaying Information Extracted by Large Language Models

Aubert-Bédouchaud, Julien, Maxime

julienaubeb@gmail.com

Nantes Université, Nantes, France; National Institute of Informatics (NII), Tokyo, Japan

Kitamoto, Asanobu

kitamoto@nii.ac.jp

ROIS-DS Center for Open Data in the Humanities, Tokyo, Japan; National Institute of Informatics (NII), Tokyo, Japan

Introduction

Recent advances in the field of natural language processing, particularly the trend towards Large Language Models (LLM) (Zhao et al., 2023), enable new possibilities to enhance the readers' experience. These technologies facilitate the extraction of meaningful information from documents, thereby improving the overall reading experience.

The objective of our system is to enhance the reading experience of e-books by swiftly identifying relevant information and presenting alternative methods for the user to engage with the content. Our system draws inspiration from hyper-reading (Sosnoski, 1999), which comprises techniques occasionally employed when interacting with digital textual content. The aim is for readers to quickly locate relevant information by employing these techniques, enabling them to engage with the content selectively.

With the internet serving as an essential platform for accessing information, reading on digital screens is increasingly popular in the education field. However, studies suggest that reading on digital screens may not be as effective as reading physical texts (Baron, 2013) (Mangen et al., 2013). This is partly due to the fact that people perceive digital reading as a means of seeking information, rather than engaging in the process of reading a full text (Baron, 2017). Implementing hyper-reading strategies could, therefore, be beneficial to e-books readers.

Our contribution is HaLLMet, an easily-customizable web application for a book reading system that enhances seamless reading experience with LLMs. The prototype accepts a standard format of e-books as input, and supports on-the-fly generation of information by LLM. It can also be

used for the evaluation of tasks tailored to modern reading habits.

Methods

Dataset

We are targeting books from the Gutenberg Project, a digital library of books, articles and essays in the public domain. Offering over 70,000 documents, the project provides access to a substantial number of important literature works in different formats, i.e. plain text, HTML page, EPUB publication or PDF document. Our prototype works with EPUB, because it is a standard e-book format supported by multiple software libraries for manipulating its data and metadata.

Large Language Model

Our prototype uses OpenAI's GPT-3.5 Turbo for rapid prototyping at a relatively low cost. To mitigate the inherent variability of generative pre-trained models, the model's temperature is set to zero, constraining output variation.

A single LLM allows multiple tasks with an instruction-tuning strategy, enabling the model to target a specific instruction. Our instruction-tuning strategy is inspired by context-faithful prompting strategies (Zhou et al., 2023) and consists of an instruction, generic examples if the output requires a specific output format and the context where the task should occur (figure 1).

Instruction: *The task the model should target.*

Example: *One or a few-shots examples, if required.*

Context: *The content that should be processed*

Answer:

Instruction-tuning strategy template

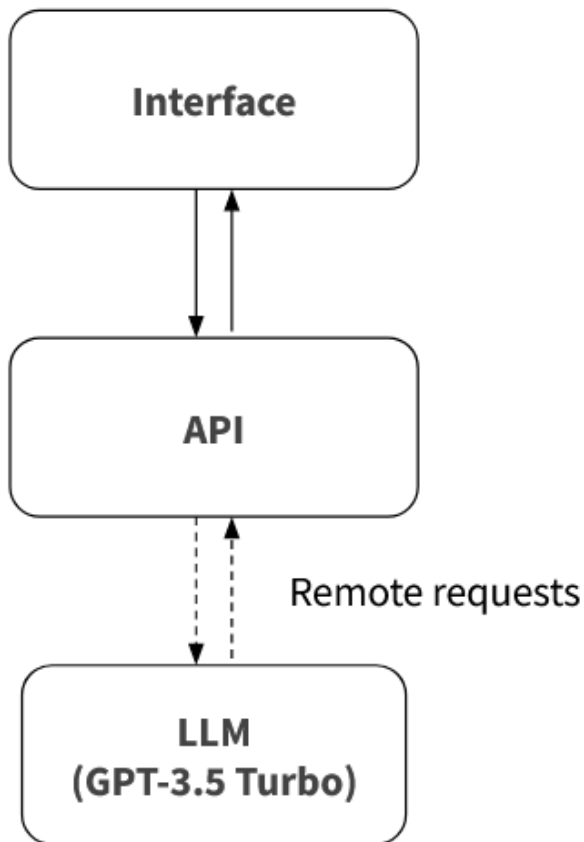
Three different kind of tasks were designed using this strategy.

- Sentence (or excerpt) extraction task: extracting batches of meaningful excerpts from the book pages.
- Sentence analysis task: generating analyses of excerpts from the surrounding context to solve the problem of lacking proper context in sentence extraction.

- Title generation task: generating a title of the situation based on the extracted content.

Web application

Our prototype uses a three-layer API structure and takes in account the needs of our project (figure 2).



Software architecture of our prototype

The interface uses web components, facilitating the management of components dependencies and allows a streamlined development flow. By adopting this paradigm, the system should ensure a cohesive and efficient environment for the design and implementation of new tasks on the user interface.

Logic and model instructions are located in the API component, whose goal is to handle all gateway operations between the user interface and the model. User interactions over the user interface request some instruction generation through the API, which will then request the model to get an output.

This structure makes it possible to modify individual layers if the project has to evolve during development and

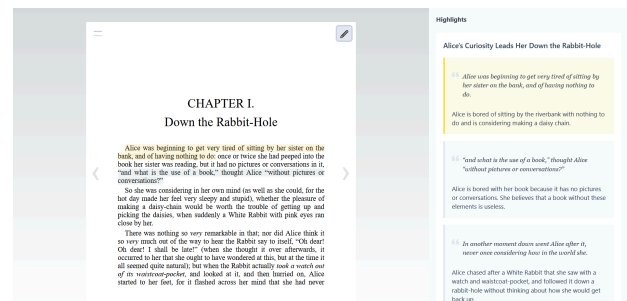
allows possibility of scaling the application when it is made available to the public. Our architecture also facilitates the integration of other models, allowing to replace GPT-3.5 Turbo with open source models.

Results

Our book-reading prototype displays the e-book content from EPUB and highlights important excerpts as shown in Figure 3. These highlights can then be further explored on a panel located on the right of the screen, allowing to navigate through book pages or highlighted excerpts. Extracted highlight are also explained by the generated title and explanation of the content to further enhance fragmentation of the information, one of the concepts of hyper-reading (Sosnoski, 1999).

This interface, showing the panel on the right, is inspired by gloss annotation. However, some users suggest that the current interface is cluttered, so we continue improving this interface based on feedback from users.

The source code of HaLLMet, including information about a demo, is available at <https://github.com/jjbes/HaLLMet>.



Interface of the web application

We also evaluated abstractive and extractive summarization of GPT-3.5 Turbo combined with our prompting strategies. We used BookSum (Kryściński et al., 2022), a dataset offering chapter summaries of literary works for Project Gutenberg books collected from various web services. Although the original dataset contains 12630 chapters, we selected only 200 randomly selected chapters due to the limitation of time and OpenAI's API requests. Figure 4 shows the result of F1 scores of common ROUGE (Lin, 2004) metrics in unigrams (R-1), bigrams (R-2) and longest common subsequence (R-L). Our model is highlighted in blue, and scores in bold indicate better instruction strategy.

BookSum-Chapter			
Model	$R-1_{f1}$	$R-2_{f1}$	$R-L_{f1}$
Heuristics			
Lead-3	14.32	2.23	8.59
Random Sentences	12.54	1.32	7.43
Extractive Oracle	42.36	9.83	20.91
Extractive Models			
CNN-LSTM	32.50	5.51	13.91
BertExt	32.06	5.37	13.68
MatchSum	30.97	5.34	13.23
HaLLMet			
HaLLMet _{Extracted Sentences}	31.40	6.15	13.49
HaLLMet _{Sentence Analysis}	28.41	8.12	12.88
HaLLMet _{Both}	23.45	7.28	10.74

Comparison of our model strategies (in blue) and other dedicated extractive models and heuristics

For extracted sentences task, the scores indicate this approach provide an amount of information similar to dedicated summarization models, indicating that the extracted excerpts are relevant for document explanation.

For sentence analysis task, contextualizing the excerpts through sentence analysis task seems to be less informative than extracted sentences. This is probably due to the nature of this task, closer to abstractive summarization.

For extracted sentences coupled with sentence analysis task, all F1 scores were lower than single tasks. This result suggests that information generated by the system is redundant. We need to find a better trade-off between the quantity and quality of information.

Conclusion

We developed HaLLMet, an easily-customizable web application designed to enhance e-book with LLMs by providing functionalities inspired by hyper-reading. While we focused the development of this application on hyper-reading strategies, the highly versatile approach could lead to other systems using HaLLMet architecture as a framework. Our system could be enhanced or modified depending on what the user needs and could be used as foundation for future works.

Bibliography

Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., Wen, J.-R. (2023) A Survey of Large Language Models.

Sosnoski, J. (1999) Hyper-readers and their Reading Engines. University Press of Colorado, pp. 161–177.

Baron, N.S. (2013) Redefining Reading: The Impact of Digital Communication Media. PMLA 128, 193–200.

Mangen, A., Walgermo, B., Brønnick, K. (2013) Reading linear texts on paper versus computer screen: Effects on reading comprehension. International Journal of Educational Research 58, 61–68.

Baron, N. S. (2017). Reading in a digital age. Phi Delta Kappan, 99(2), 15–20.

Kryściński, W., Rajani, N., Agarwal, D., Xiong, C., Radev, D. (2022) BookSum: A Collection of Datasets for Long-form Narrative Summarization.

Lin, C. Y. (2004) ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out*. Barcelona, Spain. Association for Computational Linguistics, pages 74–81

A TEI-based Approach to Data Driven Analysis of Japanese Translationese

Camilleri, Gabriele

u038475b@ecs.osaka-u.ac.jp
Osaka University, Japan

Translationese, the set of linguistic features observed more frequently in translated texts than in original texts in the same language, has long been an object of study in corpus linguistics. While initial studies focused on identifying potential candidates for translation universals (among others, Baker, 1993; Mauranen and Kujamäki, 2004; Mauranen, 2008), such as simplification or explicitation, later research has also explored the influence the source language or the target language may exert on translation choices (Chesterman, 2004). Research on Japanese translationese has suggested a higher frequency of overt personal pronouns and loanwords, abstract nouns as agents of transitive verbs, and longer paragraph length (Fukuchi Meldrum, 2009) as its chief characteristics. Japanese translations are also noted for a more "conservative" use of "role language" characteristics - stylistic variations used to convey a certain attribute of a character (Kinsui, 2003), as female characters' translated speech exhibits more frequently expressions associated with the female language stereotype (Nakamura, 2013). While the availability of large-scale Japanese parallel corpora has increased in recent years, a majority of the data is in the Japanese-English language pair (Lison and Tiedemann, 2016; Rikters et al., 2020; Morishita et al., 2022), or belongs to domains that are not particularly suitable for analyzing

the use of role language characteristics, such as scientific papers (Nakazawa et al., 2016).

This poster presents the initial stages of the design and implementation of an Italian-Japanese parallel corpus, consisting of the full texts of modern and contemporary Italian novels and their Japanese translations, for the purpose of studying translationese and role language characteristics in the underrepresented domain of literary translation, as well as in a resource-poor language pair. The methods and issues of alignment, encoding, and analysis applied to the corpus will be discussed, along with the ethical and legal concerns regarding the possibility of sharing and distributing the corpus.

Using data retrieved from the NDL (National Diet Library) Search website regarding Japanese translations of Italian novels published from 1900 to 2020 (classified as “973 Italian Literature – Fiction, Romance, Novel” under the Nippon Decimal Classification), we have selected a prominent work, *Trionfo della morte* (Triumph of death) by Gabriele D’Annunzio, which has been translated numerous times into Japanese. Eight translations from various time periods (1913; 1921; 1927; 1928; 1939; 1958; 1961; 2010) are chosen as the primary corpus for examination.

The original and its translations obtained from this sample are standardized and automatically aligned at the sentence level. To this end, the performance of two state-of-the-art alignment tools, Bleualign (Sennrich and Volk, 2010) and Vecalign (Thompson and Koehn, 2019), is evaluated on the first chapter of the novel. The alignments produced by each algorithm are compared with a reference test set of manually obtained alignments for the same sample. In this specific instance, Vecalign vastly outperforms Bleualign in terms of strict precision, recall, and F1-score, with Bleualign closely approximating the reference set at the beginning and end of the sample but losing precision in the middle section. However, as Bleualign’s output depends on the quality of the machine translation provided, further study is required.

Algorithm	P	R	F_1
Bleualign	0.40	0.38	0.39
Vecalign	0.930	0.94	0.94

Table 1: Precision (P), recall (R), and F1 for each alignment test.

The aligned texts are then converted into XML format and encoded in TEI. Using a customized Python script, the following elements, in addition to metadata and source information, are automatically annotated with their respective TEI tags: paragraphs (<p>), sentences (<s>), and dialogue and thought spans (<said>). Each speech or thought span will be then manually annotated for its speaker(s) and its addressee(s). A list of the major

characters present in the novel will also be provided for all documents, and for each character, four of their main attributes (gender, age, occupation, and socio-economic status) that have been shown to be relevant to the analysis of role language (Kinsui, 2014) will also be provided. To demonstrate the potential effectiveness of this encoding framework, quantitative pilot studies will be attempted to explore character and attribute-specific language patterns in different translations.

```
<p>
  <s>三月の午後のピンチオは夢れて見えた。</s>
  <s>折々聞こえる物音も灰色の重つた空の中へ消込んで行つた。</s>
</p>
<p>
  <said who="#Giorgio" towhom="#Ippolita">
    『
      <s>依然然うだ。</s>
      <s>自殺だ。</s>
    』
  </said>
  <s>ジョルジオが言つた。</s>
</p>
```

Figure 1: An example of the encoding in the 1913 translation.

Bibliography

- Baker, M.** (1993). “Corpus Linguistics and Translation Studies: Implications and Applications.” In Baker, M. et al (eds.), *Text and Technology: In Honour of John Sinclair*. John Benjamins, pp. 233–250.
- Chesterman, A.** (2004). “Beyond the particular”. In Mauranen, A., and Kujamäki P. (eds.), *Translation universals: Do they exist?*. John Benjamins, pp. 33–50.
- D’Annunzio, G.** (1913). *Trionfo della morte* [Triumph of death]. Translated by G. Ishikawa. Tokyo: Dainippon Toshō.
- D’Annunzio, G.** (1921). *Trionfo della morte* [Triumph of death]. Translated by O. Mikami. Tokyo: Tokasha.
- D’Annunzio, G.** (1927). *Trionfo della morte* [Triumph of death]. Translated by T. Yaguchi. Tokyo: Shiobunkaku.
- D’Annunzio, G.** (1928). *Trionfo della morte* [Triumph of death]. Translated by C. Ikuta. Tokyo: Shinchosha.
- D’Annunzio, G.** (1939). *Trionfo della morte* [Triumph of death]. Translated by G. Harada. Tokyo: Kaizosha.
- D’Annunzio, G.** (1958). *Trionfo della morte* [Triumph of death]. Translated by J. Iwasaki. Tokyo: Kawade Shobo Shinsha.
- D’Annunzio, G.** (1961–1963). *Trionfo della morte* [Triumph of death]. Translated by S. Nogami. Tokyo: Iwanami Shoten.
- D’Annunzio, G.** (2010). *Trionfo della morte* [Triumph of death]. Translated by I. Waki. Kyoto: Shuraisha.

- Fukuchi Meldrum, Y.** (2009). “Translationese in Japanese Literary Translation.” *Traduction, Terminologie, Redaction*, 22 (1): 93-118.
- Kinsui, S.** (2003). *Vācharu Nihongo: Yakuwarigo no Nazo* [Virtual Japanese: The Mystery of Role Language]. Tokyo: Iwanami Shoten.
- Kinsui, S.** (2014). *Yakuwarigo no Shōjiten* [A Glossary of Role Language]. Tokyo: Kenkyusha.
- Lison, P., and Tiedemann, J.** (2016). “Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles”. *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), pp. 923-929.
- Sennrich R., and Volk, M.** (2010). “MT-based Sentence Alignment for OCR-generated Parallel Texts”. *Proceedings of AMTA 2010*.
- Mauranen, A., and Kujamäki, P.** (2004). *Translation universals: Do they exist?*. Amsterdam: John Benjamins.
- Mauranen, A.** (2008). “Universals tendencies in translation”. In Anderman, G., and Rogers, M. (eds.), *Incorporating Corpora: the linguist and the translator*. Multilingual Matters, pp. 32–48.
- Morishita, M., Chousa, K., Suzuki, J., and Nagata, M.** (2022). “JParaCrawl v3. 0: A Large-scale English-Japanese Parallel Corpus”. *arXiv:2202.12607*.
- Nakamura, M.** (2013). *Honyaku ga Tsukuru Nihongo: Hiron wa “Onna Kotoba” o Hanashi-tsuzukeru* [Japanese Created through Translation: The Heroine Keeps on Using “Female Language”]. Tokyo: Hakutakusha.
- Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S., and Isahara, H.** (2016). “ASPEC: Asian scientific paper excerpt corpus”. *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC2016), pp. 2204–2208.
- Rikters, M., Ri, R., Li, T., and Nakazawa, T.** (2020). “Document-aligned Japanese-English conversation parallel corpus”. *arXiv:2012.06143*.
- Thompson, B., and Koehn, P.** (2019). “Vecalign: Improved Sentence Alignment in Linear Time and Space”. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP), pp. 1342–1348.

Near-synonym noun-noun patterns in the Hachidaishu Dataset

Chen, Xudong

xchen@shs.ens.titech.ac.jp
Tokyo Institute of Technology, Japan

Hodošček, Bor

hodoscek.bor.hmt@osaka-u.ac.jp
Osaka University, Japan

Yamamoto, Hilofumi

yamagen@ila.titech.ac.jp
Tokyo Institute of Technology, Japan

Overall objectives

This study is part of an ongoing project that aims to explore how to describe the extra-linguistic characteristics of lexemes in classical literary languages within the sociolectometry framework (Geeraerts et al., 1999; Speelman et al., 2003), whose aim is to measure the distances between lects. When a lexical variable contributes to distinguishing lects, we can infer the characteristics of each variant based on their biased usage choices in different lects.

As a starting point, the current study focuses on the diachronic lexical variation in the poetic vocabulary in the *Hachidaishu*. We will demonstrate using the lexical resources provided in the Hachidaishu Dataset (Hodošček and Yamamoto, 2022) to find conceptually related noun-noun (poly)lexical patterns (henceforth NN pattern) as triggers to search for potential lexical variables, such as simplex/compound variants that verbalize the same entity.

Motivations

Excluding thematic factors

Previous research on historical Japanese language variation has focused on comparing frequent or salient lexemes among different lects to infer lectal characteristics. However, the research cannot control thematic bias in the corpora. The lexical sociolectometry framework based on lexical variables can avoid thematic bias (Speelman et al., 2003: 325).

Sampling the lexical variables

Lexical studies in the Japanese linguistics tradition (cf., Kabashima, 1980), like the sociolectometry framework, requires aggregation perspectives. Therefore, we need to conduct analysis on a set of lexical variables. However, subjectivity exists in creating the set to study (De Pascale, 2019: 18).

Parallel corpora can offer objective methods for obtaining lexical variables (Tanaka and Yamamoto, 2014). However, when dealing with non-parallel corpora, it becomes difficult to determine which lexemes can be categorized as the same variables. The sociolectometry framework provides a solution to this unaddressed issue in Japanese lexical studies. Researchers collect (near) synonymy as lexical variables in three ways: a) manual, b) based on lexical resources, and c) based on distributional models. Lexical resource-based selection can be a convenient starting point. However, the Hachidaishu Dataset does not directly support the extraction of near synonymy. We need to find an alternative way to utilize the lexical resource.

Including polylexical units

We include polylexical units, although lexical sociolectometry generally excludes polylexical units (De Pascale, 2019: 186–187). We have the following reasons.

Firstly, Komatsu (2003) suggests that when analyzing classical poetic Japanese, we need to pay attention to the chain of kana characters since Japanese poetry was written with the kana strings and without any kanji characters.

Secondly, including polylexical units helps in integrating certain assumptions in lexical semantics for the future works. For example, Geeraerts et al. (1994) demonstrate that 1) peripheral members in a category are often referred to by alternative terms, such as their hypernyms (pp. 172–173); 2) entrenched concepts tend to be named with simplex forms rather than compounds (p. 175). However, we may observe exceptional situations in poetic Japanese.

Materials and methods

The Hachidaishu Dataset and old WLSP annotation

The Dataset includes potential compounds with their decompositions and categorical identifiers based on the old version of Word List by Semantic Principles (WLSP; Nakano et al., 1994; figure 1). We extract NN patterns using the WLSP annotations.

group lemma token
BG-01-1630-01-0100 年 (years)
PoS field variant

Figure 1: Format of WLSP in the Hachidaishu Dataset

Extraction rules

NN patterns follows the following rules:

1. Noun.1 + noun.2 (compounds), e.g., sakura+hana, a compound consisting of sakura (cherry blossom) and hana (flower), or
2. Noun.1 + genitive case + noun.2, e.g., sakura+no+hana, consisting of sakura, no (genitive case particle) and hana.
3. The longest common substring between WLSP metacodes of noun.1 and noun.2 should be greater than eight (to ensure two components are in the same semantic category).
4. Pattern frequency should be greater than five.

Manual selection

In each NN pattern, one of the components should ideally be conceptually interchangeable with the entire pattern. In such patterns, the main component implies the other component or the entire pattern. The rule-based process can only guarantee that two components belong to similar categories. Therefore, we conduct manual verification.

Basic statistics

For each extracted NN pattern, we count the frequency of the main component and the pattern itself to examine how poets choose between using compounds (NN pattern) or simplex forms (main components) to verbalize the same entities.

Results

We obtained a set of 93 NN patterns, among which 44 patterns, along with their main components, can be considered as potential lexical variables. This suggests the WLSP-based search is open to improvement.

We examined the variation in lexical choices between NN patterns and their main components. Currently, CHERRY is the only oncept that demonstrates a transition from a compound to a simplex form (from さくらばな to さくら) (figure 2). This indicated that CHERRY is a special concept in poetic Japanese. Although it is a frequently used object and a dominant member in the flora family, poets used compound forms rather than simplex form to verbalize it in the first seven periods. In the *Kokinshu*, the

first anthology of the *Hachidaishu*, poets also used the hypernym hana to verbalize CHERRY.

ORANGE (たちばな/はなたちばな) and PAMPAS GRASS (すすき/はなすすき) are also dominant concepts in Japanese poetry but are predominantly named using compound forms. However, unlike CHERRY, these two concepts did not demonstrate shifts in lexical choices over time.

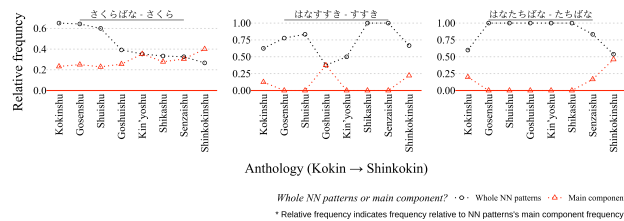


Figure 2: Exceptional lexical variables and diachronic change of variant choices

Discussion

The rule-based search only resulted limited parts of lexical choice in the vocabulary. More exceptional variables like CHERRY may remain invisible.

The NN pattern is not the only instance of near synonymy lexical variables (usually, NN patterns are hyponyms of their main components but not always). In many cases, we cannot discern the specific object when it is indicated by its hypernym. We may try domain adaptation of pre-trained language models and using the current results as training data to broaden the scope of investigation. Based on the expanded list of lexical variables, we will examine which variables can differentiate the time periods of the *Hachidaishu* and conduct statistical modeling of the lexical choices.

In the preliminary descriptive analysis, we highlighted exceptional cases in poetic Japanese that may deviate from the rules regarding simplex/compound form choices (cf., Geeraerts et al., 1994). These results suggest the importance of incorporating these polylexical aspects into further statistical modeling.

Conclusion

This study focused on rule-based searching for (poly)lexical variables from the *Hachidaishu* Dataset as a preliminary step in analyzing lexical variables. We demonstrated how lexical resources like WLSP could help the semi-automatic selection, which shows that the work of WLSP for historical Japanese (Asahara et al., 2022) will

be essential for future studies. As a result, although the rule-based search is open to improvement, the preliminary analysis showed that using near-synonymous NN patterns to integrate polylexical units into lexical variables can be significant for the poetic vocabulary.

Bibliography

- Asahara, M., Ikegami, N., Suzuki, T., Ichimura, T., Kondo, A., Kato, S. and Yamazaki, M. (2022). CHJ-WLSP: Annotation of 'Word List by Semantic Principles' Labels for the Corpus of Historical Japanese. *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*. Marseille, France: European Language Resources Association, pp. 31–37.
- De Pascale, S. (2019). Token-based vector space models as semantic control in lexical sociolectometry Leuven: KU Leuven PhD dissertation.
- Geeraerts, D., Grondelaers, S. and Bakema, P. (1994). *The Structure of Lexical Variation: Meaning, Naming, and Context*. (Ed.) Dirven, R. & Langacker, R. W. *The Structure of Lexical Variation: Meaning, Naming, and Context*. (Cognitive Linguistics Research 5). Berlin: Mouton de Gruyter doi:10.1515/9783110873061.
- Geeraerts, D., Grondelaers, S. and Speelman, D. (1999). *Convergentie En Divergentie in De Nederlandse Woordenschat. Een Onderzoek Naar Kleding- En Voetbaltermen*. Amsterdam: Meertens Instituut.
- Hodošček, B. and Yamamoto, H. (2022). Development of datasets of the *Hachidaishū* and tools for the understanding of the characteristics and historical evolution of classical Japanese poetic vocabulary. *Digital Humanities 2022 Conference Abstracts*. Tokyo: The University of Tokyo, pp. 647–48.
- Kabashima, T. (1980). *Goi/Vocabulary* (Ed.) The Society For Japanese Linguistics *Kokugogaku daijiten/Dictionary of Japanese linguistics*. Tokyo: Tokyodo.
- Komatsu, H. (2003). *Kanabun No Koubun Genri/ Principle of Constructions in Kana*. Additional Version. Tokyo: Kasama shobo.
- Nakano, H., Hayashi, O., Ishi, H., Yamazaki, M., Ishii, M., Kato, Y., Miyazaki, T. and Kirioka, A. (1994). *Bunrui Goi Hyo Furoppi Ban/Word List by Semantic Principles, Floppy Disk Version*. Vol. 5. (Kokuritsu Kokugo Kenkyujo Gengo Shori Deta Shu/National Language Research Institute Language Data). Tokyo: Dainippon shoten.
- Speelman, D., Grondelaers, S. and Geeraerts, D. (2003). Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities*, 37(3): 317–37 doi:10.1023/a:1025019216574.

Tanaka M. and Yamamoto H. (2014). Konjaku monogatari shu to Uji shuui monogatari doubun setsuwa ni okeru go no taiou – go no buntai teki kachi no kijutsu/ Word Correspondences between Tales Sharing the Same Origin Contained within the Konjaku Monogatari and Uji Shui Monogatari: A Description of the Stylistic Features of Words. *Studies in the Japanese Language*, **10**(1): 16–31.

Analysis of the Appearance Pattern Tendency of “Crying Scene” and Verification for Reproducibility of Categorization

Fukumoto, Takaki

g2122057@fun.ac.jp
Future University Hakodate, Japan

Murai, Hajime

h_murai@fun.ac.jp
Future University Hakodate, Japan

Introduction

In recent years, narrative research has been conducted in various genres such as romance and battle (Murai, 2021; Saito, 2021; Shiratori, 2021) based on the idea of narrative function proposed by Propp (Propp, 1968). Among these, narrative research has been conducted by classifying narrative works that encourage users to cry, as “crying narratives” (Fukumoto, 2022; Fukumoto, 2022).

Since there is almost no data on “crying narratives,” it is difficult to classify them mechanically. Therefore, it is necessary to perform the classification manually. This study aims to extract quantitative characteristics of “crying stories” using statistical methods by converting the stories into a sequence of digital symbols. In previous studies, a category for classifying “tear-prompting approaches” that are believed to bring users to tears was created. Moreover, the relationship between “crying scenes” and narrative functions were investigated and the objectivity of scene classification as “crying scenes” was examined using a survey. However, previous studies did not investigate the appearance pattern of “crying scenes” in terms of where they appear more frequently—in the beginning, middle or end of the story. The various types of “crying scenes” have characteristics in terms of what appears in them and

how they are used. The author’s strategy for effective use of “crying scenes” has some form of pattern. For example, in works comprising several “crying scenes”, the scene having more potential for “crying” may be more likely to be placed at the climax. In addition, the classification of “crying scenes” using the “tear-prompting approaches” was conducted by a single analyst, therefore, it lacked objectivity. It is necessary to verify the objectivity of the “tear-prompting approaches.”

Extracting the appearance pattern of “crying scene”

This study investigated the pattern of occurrence of “crying scenes.” The target narratives were five contemporary Japanese entertainment works. The selection criteria were as follows: the works are within the top 30 “crying” narrative works ranked by unspecified votes on a website; (HANABISHI, 2015) they include more than five “crying scenes”; and original works of selected works are comic or novels. The 196 “crying scenes” obtained from the five works, were classified using “tear-prompting approaches.” The classification using “tear-prompting approaches” is a multi-label classification in which each “crying scene” is labeled according to the following ten categories: “Hardship,” “Praiseworthiness,” “Relief,” “Dream Positive,” “Dream Negative,” “Bond Positive,” “Bond Negative,” “Effort Positive,” “Effort Negative,” and “Farewell” by one analyst alone. For example, if a character is persecuted for some reason that brings tears to the eyes of the audience, it is classified as “Hardship.”

The narrative works were divided into small stories by scenes in which the purpose of the protagonist changed. The appearance position of each “crying scene” in the small story was expressed as a ratio that the order of the target scene was divided by the total number of scenes within the small story. Figure 1 shows an appearance position example where small story composed of 5 scenes. The frequency of occurrence was analyzed using 0.2 as the range of ranks for the position ratio. Number of each “tear-prompting approaches” appearances in small stories are shown in table1.

Consequently, few “crying scenes” appeared in the range of 0.0 to 0.2, and in most categories, the range of 0.8 to 1.0 was most frequent. The next range in frequency was 0.6 to 0.8. However, the most frequent range only for the “Hardship” category was from 0.4 to 0.6. It is believed that several works including “crying scenes” use the pattern where the persecution, misfortune, and regret of the characters was described in the middle of the story, and thereafter, the results of the “Hardship” were described in the end of the story to induce tears.

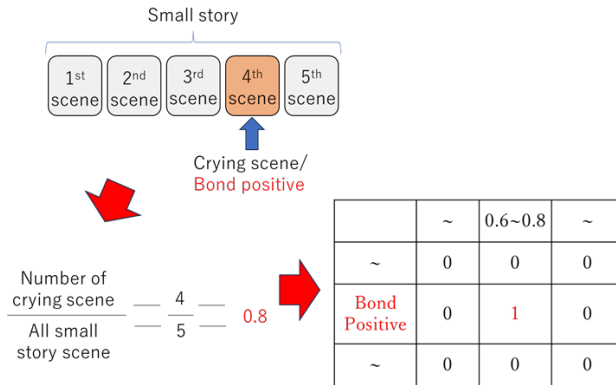


Figure 1. An example calculation appearance position of each “tear-prompting approaches”

Table 1. Number of each “tear-prompting approaches” appearances in small stories

	0~0.2	0.2~0.4	0.4~0.6	0.6~0.8	0.8~1.0
Hardship	1	5	14	7	12
Praiseworthiness	0	1	5	5	5
Relief	0	2	0	5	15
Dream Positive	0	0	2	4	9
Dream Negative	0	0	3	2	3
Bond Positive	0	2	1	13s	29
Bond Negative	2	6	2	11	17
Effort Positive	0	0	1	3	8
Effort Negative	0	1	1	2	3
Farewell	1	0	1	3	11

Verification for reproducibility of categorization

Of the 196 “crying scenes,” verification for reproducibility of categorization was performed for the top three categories—“Bond Positive,” “Bond Negative,” and “Hardship.” The sum of these three categories was 50% or more of the total data. Nine non-multi-labeled scenes were selected from the three works. These scenes were categorized in each of the three categories “Bond Positive,” “Bond Negative,” or “Hardship.”

Narrative theory researchers were asked to identify the first and second candidates of categories for each scene from the nine categories excluding “Praiseworthiness.”

In this test, “Praiseworthiness” was excluded because it appeared only when it overlapped with other categories. The results of the agreement validation revealed that the Kappa coefficient was 0.55 for the responses to the first candidate only, and 0.81 for the responses including the second candidate.

This result suggests that the objectivity of the majority three categories of “Bond Positive,” “Bond Negative,” and “Hardship” in the “tear-prompting approaches,” including the second candidate, is statistically significant.

Conclusion and future works

In this study, the appearance pattern of “crying scenes” was investigated, and a consistency examination of the classification categories was performed. The results suggest that most “crying” narrative works depict characters’ misfortunes and regret from the middle to the end of the story, and that the end of the story depicts various “crying scenes” as a result of the misfortunes and regret. In the category agreement survey, three categories, in which more than 50% of the 196 “crying scenes” were classified, were investigated using the Kappa coefficient. The Kappa coefficient for the three categories, including the second candidate, was 0.81. It suggests that the objectivity of the categories “Bond Positive,” “Bond Negative,” and “Hardship” is statistically significant.

Two issues to be addressed in the future are the limited amount of data and the unclear causal relationship between “crying scenes” and other scenes. We aim to increase the reliability of the analysis results by increasing the number of target “crying” narrative works. In addition, we will continue to investigate the causal relationship between “crying scenes” and other scenes to further clarify the “crying” narrative works.

Bibliography

Fukumoto, T., Ishikawa, K., and Murai, H., et al. (2022). Questionnaire survey on emotions evoked by “crying scenes” in stories, Computer and Humanities symposium, 2022: 227-234 (In Japanese).

Fukumoto, T., Shiratori, T. and Murai, H., et al. (2022). Classification and Pattern Extraction of Stories Evaluated as “Crying” Based on Narrative Structure Analysis, JASI2022, 1H5-OS-17b-02 (In Japanese).

HANABISHI Inc. (2015). Ranking of popular inspirational anime that make you cry! Which masterpieces do people recommend?, <https://ranking.net/rankings/best-touched-animes> (accessed 13 May 2023) (In Japanese).

Murai, H., Toyosawa, S. and Shiratori, T. et al. (2021). Dataset Construction for Cross-genre Plot Structure

Extraction, Proceedings of JADH Annual Conference 2021, pp. 93-96.

Propp, V. (1968). *Morphology of the Folk Tale*. U of Texas P, USA.

Saito, T., Yoshida, T. and Murai, H. et al. (2021). Basic Plot Structure in the Adventure and Battle Genres. Proceedings of JADH Annual Conference 2021, pp. 97-100.

Shiratori, T. and Murai, H. (2021). Historical Changes in the Typology and Characteristics of Endings in Contemporary Japanese Romance Novels. Computer and Humanities symposium, 2021: 38-43 (In Japanese).

Extracting “Darkness” in Contemporary Japanese Dark Fantasy

Kanazashi, Tomoya

g2123014@fun.ac.jp
Future University Hakodate, Japan

Murai, Hajime

h_murai@fun.ac.jp
Future University Hakodate, Japan

Introduction

Recently, there has been increasing research on the quantitative analysis and automatic generation of narrative structures from media, such as novels and comic books. However, several genres in previous studies have not been subjected to quantitative analyses or automatic generation, and dark fantasy is one of them. The term “dark fantasy” is unclear and the judgment of whether a story is a dark fantasy or not differs depending on the audience or the creator. Generally, works with numerous cruelty scenes or extreme depictions are considered dark fantasies. Considering the effects of such depictions, there are various policies such as age restrictions and warnings regarding the presence of cruel expressions. However, there are no clear criteria for judging the “darkness” of the story’s content or setting, although the “darkness” is considered to have the same effect as the descriptions. Conversely, many audiences prefer works including “darkness,” and many “dark fantasy” works are ranked top-selling entertainment in Japan today.

Purpose

In this study, modern Japanese entertainment works judged as dark fantasies by several experts in narrative research were used as target data for the analysis of dark fantasy. Scenes consisting of elements that the analyst considered “darkness” were analyzed.

To define “darkness,” we established a policy for extracting and classifying narrative elements that are judged by analyst as dark or cruel in works subjected to dark fantasy.

Subsequently, the selected narrative works were structured and analyzed. Based on the results, the elements constituting “darkness” were categorized. This study aims to quantitatively analyze the structure of dark fantasy works and extract elements of “darkness” from the plots, which would enable us to control “darkness” in the stories.

Methods

The study employs four criteria for selecting the works. First, as a clear definition of dark fantasy is lacking, the works had to be judged as dark fantasy by several experts. Second, each works should have an original comic book. This would help in increasing the validity of the narrative structure analysis by providing clear scene segmentation. Third, the work must be a dark fantasy that appears within the top-sales ranking (TORICO, 2005). This aimed to collect high-quality data. In addition, by analyzing the complete works, it is possible to analyze each section from the beginning to the end of the story. Fourth, the work must be contemporary Japanese. There is a worldwide demand for contemporary Japanese entertainment content, and many of them are among Japan’s top-ranking works. Consequently, four works were selected. Works title and authors name are shown in Table 1.

During the narrative structure analysis, the works were divided into chapters, and randomly selected for the analysis. The scenes in the selected chapters were categorized based on cross-genre narrative function categories (Murai, 2021) to enable comparisons with other genres. When a scene was considered to include some elements of darkness, an explanation of the darkness element was added to the dataset. Approximately 2000 scenes were analyzed for the four works.

Scenes including elements of darkness were manually categorized into groups based on similarity. Thus, the darkness category table includes 13 groups: “regret,” “despair,” “murder,” “suicide,” “death,” “oppression,” “mental or physical disorder,” “bullying/torture/discrimination,” “running wildly,” “abuse/slander,”

"betrayal/deception," "poverty," and "trouble." These elements were extracted using a darkness category table.

Table 1 Works title and authors

Title	Author	Scenes
Demon Slayer	Koyoharu Gotouge	547
Attack on Titan	Hajime Isayama	387
Fullmetal Alchemist	Hiromu Arakawa	385
Tokyo Ghoul	Sui Ishida	714

Results and discussion

The results exhibited that people preferred dark fantasy works containing 20–30% of the total scenes in the darkness category. Regarding the frequency of the darkness category, "murder," "mental or physical disorder," and "oppression" tended to appear more frequently. The co-occurrence relationship between the functional categories of narrative structures and darkness elements was extracted. Figure 1 shows the co-occurrence relationship of darkness categories calculated by the Jaccard coefficient. The network analysis is computed using PageRank algorithm. The combinations of "murder," "mental or physical disorder," and "oppression," "trouble," and "abuse/slander" also tended to be high. The result of an analysis using the graph structure showed that "oppression" had the highest centrality. In addition, 3-gram statistics revealed that the plot pattern of dark fantasy works resembles that of the battle genre (Murai, 2021). This suggests that the condition for dark fantasy works in modern Japanese entertainment is the presence of numerous darkness elements in a pattern similar to that of the battle genre.

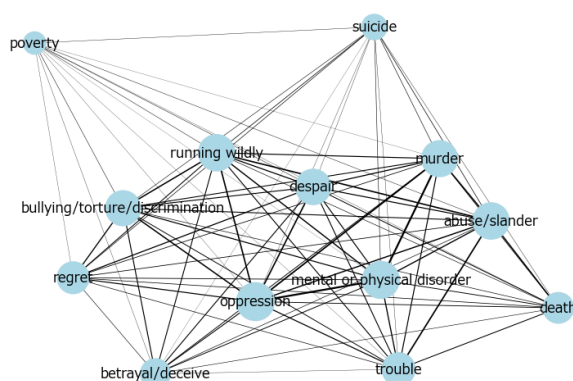


Figure 1 the co-occurrence relationship of darkness categories

Conclusion

This study attempts to clarify the narrative structure of dark fantasy genre through quantitative analysis of works judged as dark fantasy, and categorizes elements of "darkness" in the stories.

First, the darkness categories were present in 20–30% of all scenes, with "murder," "mental or physical disorder," and "oppression" as the main elements of dark fantasy stories. Second, it was observed "oppression" is central to dark fantasies.

Finally, a condition for dark fantasy works is that they contain numerous darkness elements in a pattern similar to that of the battle genre.

Future works

Future directions of this study are as follows: increasing the number of target dark fantasy works, revising the darkness category table, and comparing the results of narrative structure analysis with those of ordinary fantasy works. Regarding the objectivity of the darkness categorization, the future tasks include the reconsider of the darkness definition table and consistency validation by multiple analysts.

Bibliography

Murai, H., and Toyosawa, S., and Shiratori, T., et al., (2021). "Dataset Construction for Cross-genre Plot Structure Extraction", Proceedings of JADH Annual Conference 2021, pp. 93-96, 2021.

Murai, H., and Toyosawa, S., and Shiratori, T., et al., (2021). "Extraction of factors that characterize the structures of plot within each story genre", Information Processing Society of Japan, Computer and Humanities symposium 2021, pp. 16-23, 2021 (In Japanese).

TORICO Co.,Ltd., (2005), Manga-Volume.com: the top-sales ranking, <https://www.mangazengan.com/r/rekidai/total/> (accessed 19 October 2022) (In Japanese).

DH Research Information Portal:
A practice for "publicizing" DH
methodology

Kikuchi, Nobuhiko

kikuchi.nobuhiko@nijl.ac.jp

National Institute of Japanese Literature, Japan

The National Institute of Japanese Literature (NIJL) has been digitizing Japanese pre-modern texts both domestically and internationally, creating an integrated database called the Union Catalogue Database of Japanese Texts (formerly the Database of Pre-Modern Japanese Works). Building upon these efforts, NIJL is now advancing toward the next phase of research, aiming to establish "data-driven humanities." My presentation outlines a portal site designed to "publicize" Digital Humanities (DH) research methods, a component of the "development of humanities data analysis technology" research area within the overall plan. I also discuss the ongoing efforts to promote this portal site.

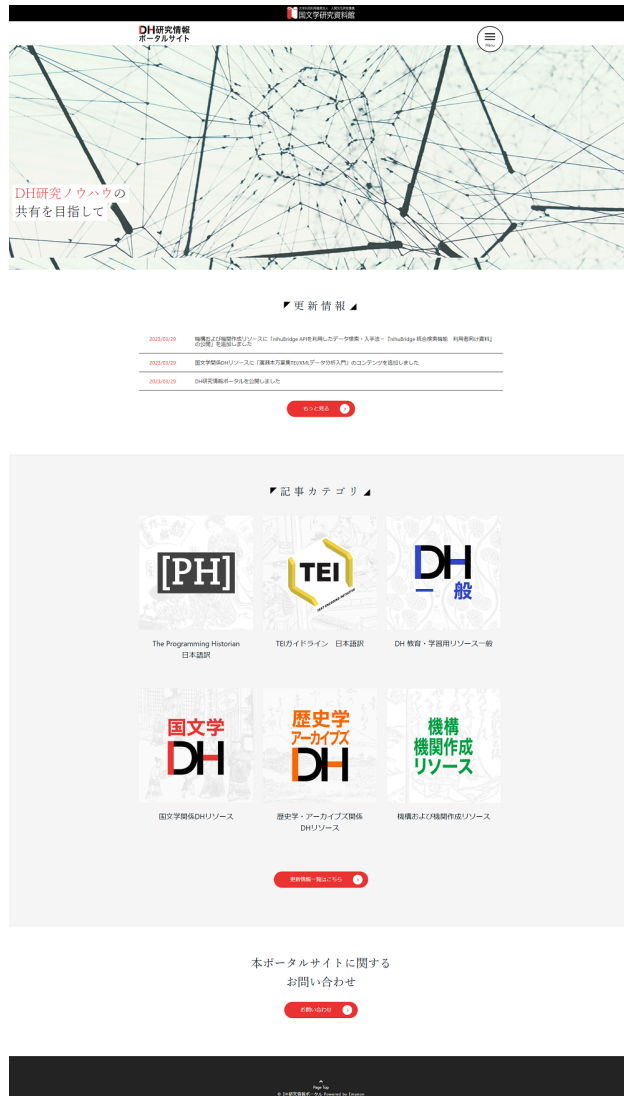
The COVID-19 pandemic has propelled the world and the educational environment to transition online. However, even before the pandemic, the development of DH training and online learning environments have been already underway, with notable progress made in other countries through platforms like Programming Historian. Focusing on East Asian studies, Vierthaler highlights DH training in Europe and the United States, summarizing DH training within Japanese studies. His article introduces the aggregation and provision of online learning materials on the Digital Humanities Japan wiki site, as carried out by Curtis et al (Vierthaler, 2020.).

In Japan, DH training has been provided, except for university education, through Kiyonori Nagasaki's TEI workshops, JADH workshops, and other initiatives. However, as Vierthaler's lack of discussion regarding inside Japan's situation paradoxically suggests, DH training and educational resource sharing in Japan have yet to overcome the language barrier. In other words, the challenge lies in organizing, consolidating, and providing online DH educational resources in Japan, while promoting information exchange between Japan and foreign countries.

By the way, in recent years, DH training programs in Europe and the U.S. have offered not only general content common to all fields of humanities, such as literature and history, but also more specialized content for each field of the humanities. Although it is challenging to demonstrate such changes statistically or diachronically, it can be assumed that once basic DH knowledge is disseminated, more specialized methods tailored to each humanities field will be required, considering both the history of the humanities itself and educational effectiveness. In other words, future DH training trends will necessitate organizing educational materials by field, or, alternatively, arranging material information from the learner's perspective.

To summarize the challenges identified earlier, we must (1) organize online DH educational resources in Japan, considering the future expansion of DH education and learning, (2) classify these resources according to each

humanities field, and (3) promote information exchange between Japan and foreign countries. I have run the "East Asian DH Portal" at Kansai University since 2020 to "import" overseas DH educational resources to Japan, for example, by translating TEI guidelines and some lessons of the Programming Historian and integrating them within East Asian studies(菊池 et. al., 2020.). With the transition of this year, 2023, I have revamped the East Asia DH Portal as the DH Research Information Portal (<https://dhportal.ac.jp/>). The DH Research Information Portal has slightly modified the categories from its predecessor in order to address issues (1) and (2). The East Asia DH Portal had been offering translations and providing information about DH tools and resources related to East Asian studies in addition to the translations of Programming Historian and TEI guidelines. Although the DH Research Information Portal has inherited the Programming Historian and TEI Guidelines categories from the East Asia DH Portal, it has been reorganized into six categories: General Educational & Learning Resources for DH, DH Resources for Literature Studies, DH Resources for Historical & Archival Studies, and resources provided by the National Institutes for the Humanities, as well as its affiliated institutions, under the category of Resources Created by the Institutes and Affiliated Entities. For instance, my colleagues and I have developed a new self-learning resource for TEI/XML data analysis of Japanese poetry materials utilizing data from the Hirose-bon Man'yoshu (菊池 et al., 2023.), and I categorized it as DH Resources for Literature Studies. For another example, the portal site introduces a document explaining methods employing the API of a database, "NIHU bridge" by the National Institute for the Humanities in the Resources Created by the Institutes and Affiliated Entities. We will continue the consolidation and provision of information related to DH training in Japan and educational resources using Japanese materials.



Nevertheless, significant challenges remain in publicizing DH methodology in Japan. The first issue concerns community building to promote information exchange between Japan and foreign countries. It is difficult for a single individual to continuously consolidate and disseminate information. Meanwhile, the current limited penetration of DH research methods means that there are few human resources available to undertake this task. Therefore, persistent efforts to expand the community of individuals who gather at the portal site are considered essential, and in doing so, it would be possible to cover a wide range of other DH methods, such as GIS, text analysis, crowdsourcing, etc., which are the primary methods of DH. The second issue revolves around understanding the impact of generative AI, such as ChatGPT, and its implications for the future. There is a risk that humanities researchers may lose the incentive to acquire expertise in different fields, as generative AI could provide them with various codes upon

request. Addressing the use of generative AI is not only a methodological concern but also a significant challenge that could impact the very raison d'être of the portal site itself. I recognize the need to address these issues to democratize and publicize DH methodology.

Bibliography

Vierthaler, P. (2020). Digital humanities and East Asian studies in 2020. *History Compass*, **18**(11): e12628 doi: 10.1111/hic3.12628. <https://onlinelibrary.wiley.com/doi/abs/10.1111/hic3.12628> (accessed 10 July 2023).

菊池信彦, 永崎研宣, 乾善彦, 海野圭介, 小川歩美 and 吉賀夏子 (2023). 和歌のXML/TEIデータ分析のための自主学習環境の構築. 研究報告人文科学とコンピュータ (CH), **2023-CH-131**(8): 1–3 <http://id.nii.ac.jp/1001/00224073/> (accessed 10 July 2023).

菊池信彦, 宮川創, ニノ宮聡 (2020). 「東アジアDHポータル」の構築と課題: デジタルヒューマニティーズの研究ノウハウのオープンな知識基盤を目指して. じんもんこん2020論文集, **2020**: 229–34 <https://cir.nii.ac.jp/crid/1050855522098879616> (accessed 10 July 2023).

Constructing fundamental behavior dataset for analysis and generation of story plots

Murai, Hajime

h_murai@fun.ac.jp

Future University Hakodate, Japan

Ohta, Shoki

g2122013@fun.ac.jp

Future University Hakodate, Japan

Ohba, Arisa

g2122015@fun.ac.jp

Future University Hakodate, Japan

Fukumoto, Takaki

g2122057@fun.ac.jp

Future University Hakodate, Japan

Aoyama, Mitsuki

g2122001@fun.ac.jp
Future University Hakodate, Japan

Okuyama, Ryogo

g2123012@fun.ac.jp
Future University Hakodate, Japan

Kanazashi, Tomoya

g2123014@fun.ac.jp
Future University Hakodate, Japan

Saito, Yuni

g2123024@fun.ac.jp
Future University Hakodate, Japan

Sato, Eiichi

g2123028@fun.ac.jp
Future University Hakodate, Japan

Tomita, Masaki

g2123041@fun.ac.jp
Future University Hakodate, Japan

Hodosawa, Tomowa

g2123055@fun.ac.jp
Future University Hakodate, Japan

Introduction

It has been clarified that it is possible to extract the common plot structure of the specific genre stories when a lot of the specific genre stories are collected (e.g. Barthes 1968, Propp 1968, and Campbell 1949). Based on those old humanistic researches, also it has been clarified that quantitative and objective extraction of those common plot structure can be executed by computational methods of recent years based on the cross-genre plot data set (Murai 2021). In those previous researches, the plot structures were described as the sequences of symbolized scenes or functions. By utilizing quantitative methods for those symbolized sequences, the differences of plot structures between genres and sub-genres have been also extracted (Murai 2022).

However, the details of the inner contents and expressions of those scenes have not been clarified yet. Therefore, it is necessary to extract characteristics of more details of expressions regarding to each plot functions and genres.

Because of fundamental model such as ChatGPT, the automatic generation of simple plots become possible in these days. If the detail characteristics of scenes can be applied in order to generate more specific long stories automatically, the result of automatic story generation would be improved. Moreover, characteristics of scene expression would be applicable for assistance of human creator.

In this research, at first common symbol sets for describing behaviors of characters' actions and describing pragmatic functions of utterances were developed. And then the behaviors of story characters were focused and frequently appearing behaviors and those patterns were extracted. Common symbols between different story genres enable to compare characteristics of each story genre. Those symbol set will be utilized for extracting common patterns for general stories. Moreover, extracted common patterns would become foundation for automatic story generation systems.

Target contents and methods

In order to compare different story genres, several popular genres in modern Japan entertainment culture were selected based on comic and game sales rankings. Selected genres were "Adventure", "Battle", "Love", "Detective", and "Horror". In order 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 protagonist of the story (Nakamura 2020).

After that, selected stories were divided into plot elements (scenes) and categories were inductively constructed manually (Murai 2022). The category table for the function of plot elements includes 29 large category and 191 small categories. For example, some scene is categorized as "disturbance" function as large category, it may be categorized also "betrayal" function or "combat" function as small category. As a result, 3185 scenes in 273 stories of 5 genres were extracted (Table 1).

Table 1 Analyzed stories and scenes

	Stories	Scenes	Average scenes
Adventure	291	2372	8.15
Battle	454	4276	9.42
Love	190	1823	9.59
Detective	268	4049	15.11
Horror	339	3878	11.44

Based on those scenes, characters' actions were categorized by utilizing the category table for actions of characters. The category table for actions of characters includes 1048 action types and about 10000 Japanese concrete vocabulary. As a result, 338 types of 4909 behaviors about actions of characters were manually categorized. In addition to those action data set, the roles and attributes of the agent and the recipient of each action were also categorized based on the table for character attributes that includes 41 large categories and 151 small categories (Table 2).

Extracted actions were aggregated in 5 genres, large and small category of functions of scenes, roles and attributes of characters and also sequential patterns of actions were extracted by n-gram.

As a whole, the most frequently appeared 5 actions were "know", "attack", "hope", "meet", and "teach". However, the result varied by genres. For instances, the most frequently appeared 5 actions in love genre were "courting", "love", "know", "help", and "dating". Those results are also dependent on the functions of scenes. For instances, in the scenes, that were categorized as "intention" function, the top 5 actions of characters changed to "wish", "request", "decision", "suggestion", and "know". Moreover, the roles and relationships between the agent and the recipient of each action affect on appeared actions. In the case that the agent is the protagonist and the recipient is the acquaintance of the protagonist, the top 5 frequently appeared actions become to "meet", "courting", "criticize", "help", and "association". On the other hand, in the case that the agent is the protagonist and the recipient is the rival of the protagonist, the top 5 become to "disturbance", "protect", "rebukey", "ridicule", and "assistance". Those differences signify the characteristics and influences of individual contexts. The frequently appearing sequences of actions also are affected on those contexts.

In addition to that, 1500 scenes were randomly extracted and utterances of characters within those extracted 1500 scenes were categorized based on the category table for pragmatic functions. The category table for pragmatic functions includes 19 categories: "introduction", "explanation", "thinking", "intention", "desire", "request", "suggestion", "question", "response", "acknowledgment", "rejection", "future", "evaluation", "reproach", "blame", "gratitude", "emotions", "solicitude", and "joke". As an example, the frequently appeared scenes and pragmatic functions at the detective genre are showed in table 3. As for the actions of characters, those categorized utterances were stored with various attributes such as roles of the speakers and listeners. Aggregated results showed that the functions of utterances changed greatly depending on the context, as in the case of actions of characters.

Table 2 Frequently appeared behaviors of characters

All	Adventure	Battle	Love	Detective	Horror
know	452	96	142	87	82
attack	188	70	95	68	36
hope	167	69	57	67	36
meet	158	62	46	60	35
teach	148	53	40	57	30
help	133	46	39	43	29
inform	113	45	35	42	27
criticize	107	37	35	40	22
request	105	35	31	24	19
courting	95	32	30	20	18

Table 3 Frequently appeared scenes and pragmatic functions at detective genre

Function of scenes	Pragmatic functions							
	introduction	explanation	thinking	request	question	response	emotions	reproach
information disclosure	18	86	45	3	24	10	9	7
correct reasoning	65	229	619	23	104	41	26	24
encounter	49	128	74	16	36	30	16	14
discover	22	39	81	5	7	0	10	11
requesting	19	46	49	20	19	16	2	7
investigation	30	67	95	14	22	13	25	15
revelation	19	66	57	6	11	3	4	4
confession	15	44	22	3	6	1	6	6
deceive	10	76	63	17	16	2	9	7
rebukey	15	32	41	13	4	4	4	10
clue	27	125	165	4	55	21	18	9
riddle	24	66	69	12	28	11	16	18
Total	323	1023	1391	139	334	153	145	135

Conclusions and future works

In order to clarify the details of story scenes, the actions and utterances of characters were categorized by utilizing various attribute tables and fundamental data set was developed. It is clarified that those actions varied depending on the various contextual information. If those characteristics can be reflected on automatically generated stories in appropriate way, the resultant texts would become more convincing and understandable.

Bibliography

- Barthes, R.,** (1968). *Elements of Semiology*. Hill and Wang, New York, USA.
- Campbell, J.,** (1949). *The Hero with a Thousand Faces*. Pantheon Books, USA.
- 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.
- Murai, H., and Toyosawa, S., and Shiratori, T., et al.** and (2021). "Dataset Construction for Cross-genre Plot Structure Extraction", *Proceedings of JADH Annual Conference 2021*, pp. 93-96, 2021.

Murai, H., and Toyosawa, S., and Shiratori, T., et al. (2022). "Extraction of Typical Story Plot Patterns from Genres within Japanese Popular Entertainment Works", ICC2022, 2022.

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).

Propp, V.,(1968). Morphology of the Folk Tale. U of Texas P, USA.

Extracting the Relationship Between the Emotions Evoked in the Story and Acoustic Features of the Music

Okuyama, Ryogo

g2123012@fun.ac.jp
Future University Hakodate, Japan

Murai, Hajime

h_murai@fun.ac.jp
Future University Hakodate, Japan

Introduction

We empathize with, or forecast, the development of a story while reading it. In particular, the music used in each scene of the movie would have the effect of arousing emotions that the author is trying to evoke through the content of each scene.

This study assumes that emotions aroused by music are related to those evoked by the content of the scene in the story. The purpose of this research was to extract the relationship between the emotions evoked when watching a story and listening to music to presume the emotions stimulated by music.

Sentiment analysis for stories has also been done so far (Tomas, 2021) (Andrew, 2016).

By clarifying the relationship between story and music, it is considered possible to support creators. Moreover, it will be useful for automatic background music (BGM) selection technology based on video composition, effect selection, and emotion level.

Material and Method

The target stories were selected based on four standards. First, stories were selected from movies because this research required both music data and the emotions evoked in the story. Second, the theme of the stories was daily life. This study assumes that it can be applied to daily life in the future. Third, human emotions (joy, anger, pathos, and humor) appear well balanced in the story since various emotions are obtained from a single story. Fourth, the story was of high quality because this guarantees analysis quality.

In this study, the movies were selected based on box office data and the Japan Academy Prize for the past 10 years from 2010 to 2019, before the COVID-19 epidemic. In this research, "Confessions" and "Umimachi diary," which satisfy the selection criteria mentioned above were analyzed.

The acoustic features of the music and emotional features of the selected story were analyzed. In the story, emotion tags were added to the scenes. We analyzed eight basic emotions proposed by Plutchik (Robert, 1980)—joy, trust, fear, surprise, sadness, disgust, anger, and expectation as emotion tag.

In addition, four acoustic features were analyzed (BPM, spectral centroid, chroma vector, and zero-crossing rate) (Okuyama, 2022). The BPM indicates the number of beats per minute and changes according to the genre. The spectral centroid indicates the center of gravity of the spectrum and changes according to the impression of sound brightness. The chroma vector indicates how much of a particular frequency is included in each scale and varies with the intensity of the notes in the scale. The zero-crossing rate indicates the frequency at which the wave changes from positive to negative, or vice versa, from the center when drawing the waveform of a song. The zero-crossing rate changes depending on how loud the song is.

Multiple regression analysis was used to extract the relationships between the four acoustic features and eight basic emotions using a combined dataset.

Result

One hundred twenty-seven scenes (ninety-two scenes from "Confessions" and thirty-five scenes from "Umimachi Diary") were analyzed. Based on these data, multiple regression analysis was performed for emotion tags and acoustic features. Results of multiple regression analysis are shown in Table 1 to Table 8. Emotion tags were the target variables, and acoustic features were the explanatory variables. As a result, the acoustic feature of the centroid was at a 5% significance level for surprise and fear. The acoustic feature of the chroma vector had a 5% significance

level for the emotion of disgust. The acoustic features of the BPM unit were marginally significant for the anticipation emotion. The acoustic features of the centroid were marginally significant for sadness. The acoustic feature of the zero-crossing rate was marginally significant for trust. However, the effect of anger could not get a statistically significant result.

Table 1. Multiple regression about “Joy”

	Estimate	Std. Error	t value	Pr(> t)
Intercept	0.6611641	0.282917	2.337	0.0211
BPM	-0.0053985	0.002157	-2.503	0.0136
Centroid	-0.000197	0.000127	-1.557	0.122
Chroma Vector	0.0272585	0.020381	1.337	0.1836
Zero Crossing Rate	0.0001459	0.00011	1.33	0.1858

Table 2. Multiple regression about “Sadness”

	Estimate	Std. Error	t value	Pr(> t)
Intercept	4.55E-01	3.57E-01	1.274	0.2051
BPM	6.23E-06	2.72E-03	0.002	0.9982
Centroid	2.73E-04	1.60E-04	1.711	0.0897
Chroma Vector	-2.75E-02	2.57E-02	-1.068	0.2877
Zero Crossing Rate	-1.86E-04	1.38E-04	-1.34	0.1827

Table 3. Multiple regression about “Trust”

	Estimate	Std. Error	t value	Pr(> t)
Intercept	-6.46E-02	2.26E-01	-0.286	0.7754
BPM	7.43E-04	1.72E-03	0.432	0.6668
Centroid	-1.22E-04	1.01E-04	-1.205	0.2306
Chroma Vector	2.36E-03	1.63E-02	0.145	0.885
Zero Crossing Rate	1.47E-04	8.75E-05	1.674	0.0967

Table 4. Multiple regression about “Fear”

	Estimate	Std. Error	t value	Pr(> t)
Intercept	-0.4216926	0.260565	-1.618	0.1082
BPM	0.0034561	0.001987	1.74	0.0844
Centroid	0.0002768	0.000117	2.375	0.0191
Chroma Vector	-0.0012376	0.018771	-0.066	0.9475
Zero Crossing Rate	-0.0001495	0.000101	-1.481	0.1412

Table 5. Multiple regression about “Surprise”

	Estimate	Std. Error	t value	Pr(> t)
Intercept	-0.2149924	0.301913	-0.712	0.4778
BPM	0.0015621	0.002302	0.679	0.4987
Centroid	0.0002682	0.000135	1.986	0.0493
Chroma Vector	0.0205722	0.02175	0.946	0.3461
Zero Crossing Rate	-0.0002174	0.000117	-1.859	0.0655

Table 6. Multiple regression about “Disgust”

	Estimate	Std. Error	t value	Pr(> t)
Intercept	3.09E-01	3.06E-01	1.009	0.3151
BPM	2.68E-03	2.34E-03	1.149	0.2527
Centroid	2.74E-05	1.37E-04	0.2	0.8419
Chroma Vector	-5.54E-02	2.21E-02	-2.511	0.0134
Zero Crossing Rate	-8.14E-06	1.19E-04	-0.069	0.9454

Table 7. Multiple regression about “Anger”

	Estimate	Std. Error	t value	Pr(> t)
Intercept	2.40E-01	1.65E-01	1.458	0.148
BPM	-2.18E-04	1.26E-03	-0.174	0.862
Centroid	-2.12E-05	7.37E-05	-0.288	0.774
Chroma Vector	-1.67E-02	1.19E-02	-1.403	0.163
Zero Crossing Rate	-6.18E-07	6.38E-05	-0.01	0.992

Table 8. Multiple regression about “Expectation”

	Estimate	Std. Error	t value	Pr(> t)
Intercept	-1.94E-01	2.63E-01	-0.737	0.4624
BPM	3.82E-03	2.01E-03	1.901	0.0596
Centroid	-9.88E-05	1.18E-04	-0.84	0.4028
Chroma Vector	-9.56E-03	1.90E-02	-0.504	0.615
Zero Crossing Rate	4.78E-05	1.02E-04	0.469	0.64

Discussion

From the results, it is considered that the emotion of joy can be estimated using the BPM. In addition, it is thought that the emotions of fear and surprise can be estimated by the centroid, and the emotions of disgust can be estimated by the chroma vector.

Expectation sentiment was not statistically significant in the BPM unit, but the p-value was less than 10%. Therefore, the BPM unit can be used as a reference value to estimate the emotions of expectations. Similarly, it is possible to use the centroid for the emotion of sadness and the zero-crossing rate for the emotion of trust as reference values for emotion estimation.

Anger was not associated with any acoustic features. There are two possible reasons for this. First, data on anger are lacking. Approximately 20 data points were obtained for other emotions, but only six data points were obtained for anger. Therefore, it is conceivable that no significant results would be obtained. Second, there was no relationship between anger and the four acoustic features used in this study. Therefore, it is possible that the emotion of anger is related to the features used in this study or that there is no relationship between the emotion of anger and acoustic features.

Future works

The amount of data was small because only two stories were analyzed. We plan to increase the number of target stories for the analysis to improve the quality of the results. Some emotions showed a 5% significance level for the analyzed acoustic features. However, some emotions had a marginally significant relationship with acoustic features, and anger had no relationship with any of the four acoustic features. Therefore, it would be useful to investigate the relationship between emotions and other acoustic features.

Bibliography

- Andrew, J. R., and Lewis, M., and Dilan, K., et al.,** (2016). "The emotional arcs of stories are dominated by six basic shapes", *EPJ Data Sci*, 5(1).
- Okuyama, R., and Kanazashi, T., and Shirakawa, R., et al.,** (2022). "Construction of an automatically generated game system to assist in the creation of game scenarios, sound effects, and background images", *The Japanese Society for Artificial Intelligence*, 2022 (In Japanese).
- Robert, P.,** (1980). "A general psychoevolutionary theory of emotion", *Theories of Emotion*, pp.3-33, Elsevier, Netherland.
- Thomas, S., and Katrin, D., and Christian, W.,** (2021). "Emotion Classification in German Plays with Transformer-based Language Models Pretrained on Historical and Contemporary Language", *Association for Computational Linguistics*, pp. 67-79.

Online Reaction towards ChatGPT Ban from Education

Takagi, Miu Nicole

miu.n.takagi@toki.waseda.jp
Waseda University, Japan

Ohman, Emily

ohman@waseda.jp
Waseda University, Japan

Introduction

ChatGPT was released on November 30th, 2022. Only a few days later, it disrupted the evaluation of education and how students interact with their assignments. Without any reliable tools available to evaluate which submissions were created with ChatGPT, New York City (NYC) public schools banned the use of the AI tool by not only students but teachers as well.

In this study, topic modeling will be used to explore online reactions to the NYC ban. Reactions have mostly been skeptical of the effectiveness of preventing plagiarism and instead called for a review of the education system. Meanwhile, when ChatGPT was banned on the question-and-answer website for programmers, StackOverflow, reactions seemed more agreeable with the ban, emphasizing its generation of incorrect answers.

Data and Method

Three recent Reddit threads, Reddit (2023c), Reddit (2023b), and Reddit (2023a) were selected for data collection due to them having the most upvotes about their respective topics, with 28.9k for NYC bans, 6.6k for one StackOverflow thread, and 1.5k votes for another StackOverflow thread, at the point of data collection.

A total of 3958 comments were collected from Reddit with regards to ChatGPT's ban from two different places; New York City public schools and StackOverflow. 2663 comments belong to the thread on bans by NYC public schools, while 856 belong to the StackOverflow thread with 6.6k upvotes, and 439 belong to the StackOverflow thread with 1.5k upvotes.

Data were first preprocessed. Each word was then tokenized by Gensim's `simple_preprocess`, and a word cloud was generated for each thread and for the combination of StackOverflow threads to identify keywords that should be included in the list of stop words brought from NLTK. Additional stop words include, "will," "ChatGPT," and in the case of StackOverflow threads, "StackOverflow."

Following preprocessing, topic modeling was conducted using LDA through Gensim. To evaluate the number of topics that should be created by the model, coherence scores against the number of topics were calculated for each Reddit thread, as shown in Figure 1. Topic numbers that formed the highest peak, but still remained less than 100, were selected

for each thread; 38 topics for NYC thread, 80 for the larger StackOverflow thread, 38 for the smaller StackOverflow thread, and 23 for the combined dataset for StackOverflow.



Figure 1: Coherence Scores

Following the evaluation, bigrams, trigrams, and co-occurrence networks were visualized for each thread utilizing nplot, an analysis and visualization module for natural language processing created by (takapy0210, 2022).

Results

New York City Ban

There seemed to be skepticism against the effectiveness of the ban of ChatGPT from public schools in New York City. Top keywords in the thread were related to students' critical thinking and problem-solving skills, how students can bypass the school firewall, and the education system. Similar topics could also be observed in 2b. In the trigram, specific methods, such as the utilization of virtual private networks and proxy servers - [bypass, school, firewall], [virtual, private, network], [private, network, vpn] - can be observed. Additionally, concern for cheating with ChatGPT could also be observed, though it was lower ranked than the aforementioned topics.

In the co-occurrence network, 3, central keywords were "tool", "thinking", "like", "people", "way", "work", "ai", "school", and "Wikipedia". Outside the central cluster, words related to education, internet access, and firewall were observed.

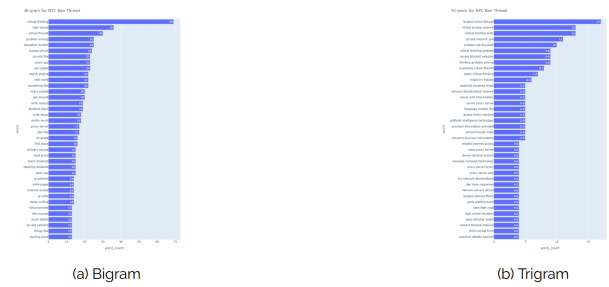


Figure 2: N-grams of Reaction to NYC Ban

Co-occurrence network for NYC Thread

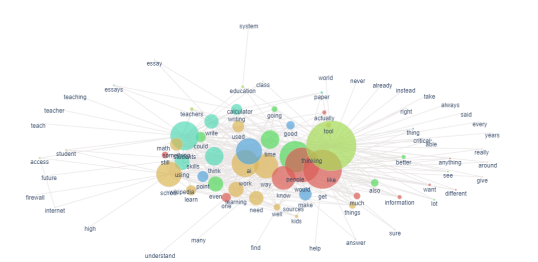


Figure 3: Co-occurrence network of keywords in NYC ban Thread

StackOverflow Ban

With regards to the StackOverflow-ban thread, in 4a, top keywords were regarding the Turing test, training set or data of ChatGPT, and wrong answers provided by the chatbot. For the trigram, however, topics were more varied, though common topics were ["incorrect", "lot", "cases"], ["gets", "stuff", "wrong"], and ["scary", "confidently", "incorrect"]. In the co-occurrence network, central keywords were "like", "people", "ai", "answers", and "correct", similar to the NYC ban thread.

In the cooccurrence network, 5, results were similar to the other thread for central key terms. However, no major thematic clusters were observed outside the central area.

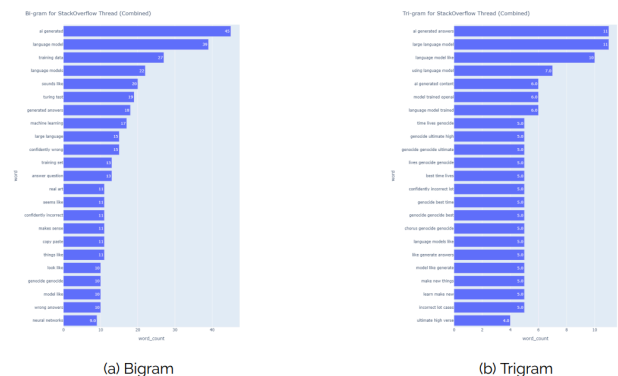


Figure 4: N-grams of Reaction to StackOverflow Ban (Combined)

Co-occurrence network for StackOverflow Thread (Combined)

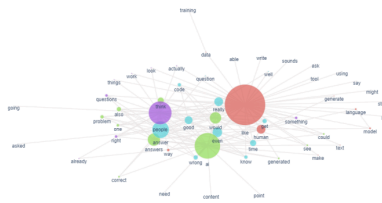


Figure 5: Co-occurrence network of keywords in StackOverflow ban Thread (Combined)

Conclusion and Discussion

Reactions against the bans in NYC and StackOverflow can be observed to be somewhat different. While NYC focused more on education, StackOverflow's discussion was more regarding the training dataset of ChatGPT. This difference can be explained by the threads' topics and audiences being different, though both belonged to technology-related threads.

Attitudes toward the ban were also seemingly different. Redditors commenting on the NYC thread were skeptical, raising examples of how students could circumvent the ban. There seemed to be suggestions to take the situation as an opportunity to develop students' critical thinking skills to allow them to not have to rely on AI such as ChatGPT. Supporting this, Rudolph et al. (2023) suggested that AI such as ChatGPT should be incorporated into an environment where students are invested in their own learning.

StackOverflow's discussion was more technical. Support was shown for the ban due to the AI not being the most accurate in providing quality code. Redditors on this thread seemed to generally be more supportive of the ban, concerned with the accuracy of text generated by the AI, as found in the results section. As noted by Chatterjee and Dethlefs (2023), the lack of accuracy for certain topics is due to the model being trained on open-domain data available on the internet, which is known to not always provide the most accurate or correct information. In this sense, in areas such as forums, where the most accurate answer is desired, it may be better to ban the use of such AI as the receiver of the answer may not always be aware that AI was used to generate the answer.

In the past, education systems reacted negatively towards the use of Wikipedia, and further back, the use of calculators. Today, they are now actively used in classrooms as educational tools. As AI becomes more mainstream and readily available, instead of reacting in extremities through bans, Redditors seem to instead want them to be educational opportunities to develop digital literacy. Yet, the banning of AI in question-and-answer forums, such as StackOverflow,

seems to be viewed as beneficial. As an online environment where users can receive help on problems, the use of AI, when there is a risk of inaccuracy, was deemed detrimental to the user's educational experience.

Bibliography

Chatterjee, J. and Dethlefs, N. (2022). This new conversational AI model can be your friend, philosopher, and guide ... and even your worst enemy. *Patterns*, 4(1):100676, 2023. <https://www.sciencedirect.com/science/article/pii/S2666389922003233> (accessed 28 January 2023).

Reddit. (2023a). ChatGPT AI Generated Answers Banned On Stack Overflow, https://www.reddit.com/r/programming/comments/zd71vl/chatgpt_ai_generated_answers_banned_on_stack/ (accessed 28 January 2023).

Reddit. (2023b). StackOverflow to ban ChatGPT generated answers with possibly immediate suspensions of up to 30 days to users without prior notice or warning, URL https://www.reddit.com/r/programming/comments/zhpkk1/stackoverflow_to_ban_chatgpt_generated_answers/ (accessed 28 January 2023).

Reddit. (2023c). NYC Bans Students and Teachers from Using ChatGPT | The machine learning chatbot is inaccessible on school networks and devices, due to concerns about negative impacts on student learning, a spokesperson said., https://www.reddit.com/r/technology/comments/103gran/nyc_bans_students_and_teachers_from_using_chatgpt/ (accessed 28 January 2023).

Rudolph, J., Tan, S. and Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, 6(1)

takapy0210. (2022). nlplot: Analysis and visualization module for Natural Language Processing, <https://github.com/takapy0210/nlplot> (accessed 28 January 2023).

Development of a dataset for comparison between predicate verb phrases in the Kokinshu and their contemporary translations

Yamamoto, Hilofumi

yamamoto.h.al@m.titech.ac.jp
Tokyo Institute of Technology, Japan

Hodoscek, Bor

hodoscek.bor.hmt@osaka-u.ac.jp
Osaka University, Japan

Chen, Xudong

chen.x.aj@m.titech.ac.jp
Tokyo Institute of Technology, Japan

1 Introduction

To compare the chain structure of predicate verb phrases in the Kokinshu (ca. 905) with contemporary Japanese translations, we developed a parallel dataset comprising ten modern translations. We then examined the differences in verb phrases between the original classical Japanese poems and their contemporary Japanese translations.

As predicates serve as both condensed and essential components of texts, we hypothesize that predicates in the original poems (OP) may contain diverse meanings due to the 31-syllable restriction of the OP format. We attempted to classify all elements in the contemporary translations (CT) into syntactic, contextual, and other elements. For simplicity, we only analyzed poems with verbs in plain form. (Suzuki 1965)

2 Methods

We used alignment techniques to subtract verbs in the OP from their aligned predicates in the CT, and subsequently classify the remaining elements of the predicates of CT into syntactically and/or contextually added elements. For the Kokinshu, we used vocabulary data for 1,000 poems from the Kokinshu collection, excluding those not in the 5/7/5/7/7 form, such as *nagauta* and *sedoka*, from the Zenodo Hachidaishu vocabulary dataset (Yamamoto and Hodošček 2021). The poem texts (OP) in this dataset are from the Nijuichidaishu database compiled by the National Institute of Japanese Literature, available from the Center for Open Data in the Humanities/CODH.

On the other hand, the 10 contemporary translations (CT) corresponding to each Kokinshu poem were manually typed using the contemporary translations from commercially available annotated books, tokenized into smaller units, and saved in a format corresponding to the OP dataset. Annotations were taken from sources such as the Shin-Nihon Koten Bungaku Taikei Bon Nijuichidaishu (New Japanese Classical Literature Compendium) (Kojima and Arai 1989, Katagiri 1990, Komachiya 1990, Kubota and Hirata 1994, Kawamura et al. 1989, Katano and Matsuno 1993), Shincho Nihon Koten Shusei, and Shin-Kokinshu (Kubota 1979), among others.

We will examine the following points concerning verbs: 1) whether the same verb is used in OP and CT; 2) whether syntactical elements are added to CT in order to maintain a meaning equivalent to that of OP; and 3) whether there

are any other elements added to CT based on the context. Using the alignment techniques, we will subtract the verb of OP from the predicate of CT, then classify the remaining elements of the predicate of CT into syntactically and/or contextually added elements.

According to the distinctions of classical Japanese (CJ) verbs shown in Table 1 (Yamamoto 2005), the plain form of CJ verbs will be translated into the four ways: i.e., the first pattern is that a verb of CJ (Voj) will be simply replaced with the historically equivalent or literally equivalent form of the verb of modern Japanese (MJ; Vmj) such as *nagaru* → *nagareru*; the second pattern is that Voj will be replaced with Vmjand *-teiru* added to it such as *nagaru* → *nagareteiru*; the third pattern is that Voj will be replaced with Vmjand *-ta* attached to it such as *nagaru* → *nagareta*; and the fourth pattern is that Voj will be simply replaced with Vmj, or after being replaced, auxiliary verb, *-u/-daro*, which indicates conjecture is attached to it such as *nagaru* → *nagareru(daro)*.

The problematic cases are the second and third patterns. Takahashi (1983) states that verbs of MJ also can express the same fact in both past and non-past. We should pay special attention to these types of verbs and must judge their tense and aspect based on their contexts. The plain form of CJ verbs, therefore, cannot be simply replaced with that of MJ verbs (Kato 1986: 62).

Table 1: Patterns of conversions from verbs of classical Japanese (CJ) to those of modern Japanese (MJ) based on the description in Yamamoto (2005); *V_{oj}* and *V_{mj}* indicates a verb of CJ and the corresponding verb with MJ respectively; *nagaru* (flow) of CJ is assumed as the same as *nagareru* (flow) of MJ.

	rules		examples	
	CJ	→ MJ	CJ	→ MJ
1	<i>V_{oj}</i>	<i>V_{mj}</i>	<i>nagaru</i>	<i>nagareru</i>
2	<i>V_{oj}</i>	<i>V_{mj}-teiru</i>	<i>nagaru</i>	<i>nagareteiru</i>
3	<i>V_{oj}</i>	<i>V_{mj}-ta</i>	<i>nagaru</i>	<i>nagareta</i>
4	<i>V_{oj}</i>	<i>V_{mj}(-u)</i>	<i>nagaru</i>	<i>nagareru(darō/dearō)</i>

3 Results

We identified four categories of patterns: 1) an obsolete verb is simply superseded with the verb currently used in MJ; 2) When compound verb elements are added to the original verb in order to strictly convey the nuances of the ancient Japanese verb; 3) a single verb is replaced with a verb phrase due to nuances between Classical Japanese (CJ) and MJ; and 4) a single verb is replaced with a verb phrase because the predicate contains a pun. Due to space limitations, Table 2 shows examples of ten different modern translations of phrases ending in ‘*nagareru*’ only.

Examples of the first category include obsolete verbs such as *tagitsu* (seethe) in KKS 830 and *utsurou* in KKS 232, which are replaced with MJ equivalents. *Tagitsu* is replaced with *sakamaku* (roll, boil, or rage), *hageshiku* *nagareru* (flow rapidly), or *nagareochiru* (flow down). The second type is patterns using compound verbs to express meanings precisely, such as *teru* (shine) and *sou* (go along

with). In many cases, *teru*(shine) is used with *haeru*(shine, glow) or *kagayaku*(shine, glitter). On the other hand, *sou*(go along with) basically means ‘add’. Some CT include *nagareru*(flow), such as in *nagarekuwawaru*(flow and add) and *nagarekomu*(flow into), which are derived from their context.

In the third category verbs like *tanomu*in KKS 555 and KKS 773 are modified by using *-suru*such as in ‘*tanomi ni suru*’ or a causative form such as *-saseru*. The expression, ‘*tanomi ni suru*’ (expect) or ‘*tanomi ni omowaseru*’ (expect) are used in CT, since *tanomu* in MJ generally means ‘ask for something’, but the sense of *tanomu*(expect/wish) as used in CJ is not used in MJ.

The fourth and most complex category involves verbs containing puns such as *kaku*in KKS 761. Here, it is difficult to find common elements between the translations. In this instance, *kaku*cannot only mean ‘write’ but can also mean *kazu-kaku*(count numbers), *mo-gaku*(suffer), or *ha-gaku*(the action of a bird, i.e., it plumes itself; *hameans* ‘wing’ or ‘feather’).

4 Discussion

The most important point to be mentioned in this analysis is that there are not many elements newly added to CT, despite the various patterns of the predicates of CT. Although this is the most unexpected result, at the same time, it suggests that the translators strictly employed a word for word method in their work. In fact, the variation in predicates can be mainly attributed to syntactical distinctions between CJ and CT. In order to further explain this point, we discuss three relevant issues: unmatched verbs, suffix selection, and newly added elements. Replacing an OP verb in with another verb is the most apparent method for adding previously unwritten elements to the CT. Due to space limitations, only the example *nagaru* (flow) as in Figure 1 is shown by the automatic aggregation of the elements added to the modern translation.

First, when the verb of OP is obsolete, a contextually similar verb is used: e.g., *tagitsu* (seethe) is replaced with *sakamaku* (roll, boil, or rage), or *hageshiku nagareru* (flow rapidly); *utsurou* (transfer) is replaced with *iroaseru*(fade) or *otoroeru*(decline; become weak).

Second, to elaborate or clarify meaning, a verb of OP is replaced by a compound verb in CT that incorporates the OP verb as a part of the compound verb. Compound verbs add more information to the OP, such as OP *teru* (shine) → CT *terikagayaku*(shine, glitter). In case of *omou*(think), the meaning of spontaneity can be emphasized by changing the form of the verb, such as *omou* → *omowareru*, “*omoi ga suru*,” “*omoi o haseru*.”

Lastly, when the OP verb comprises a pun, its translation becomes complex, as seen with *kaku* (write) in KKS 761. A CT sentence with a pun is longer than one without, as it includes two or more meanings implied by the pun. All

predicate patterns are considered modifications of the OP verb, made through translators’ interpretations based on context. This modification is crucial for the appearance of non- literal elements of OP in the CT predicates. However, in addition to mismatched elements, certain aspects remained unclarified, such as the use of suffixes, elements newly added to the modern language, and elements corresponding to the auxiliary verb, *ramu*.

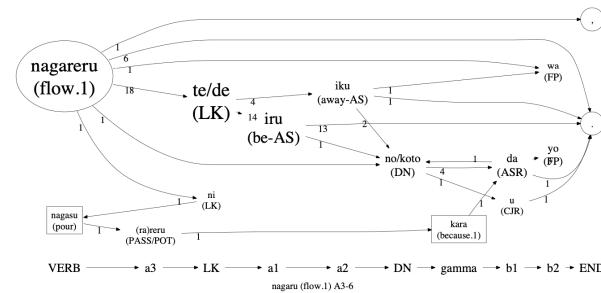


Fig. 1: Construction of the predicate of *nagaru* (flow): *nagaru* (flow.1) appears in 3 poems; it appears in KKS 284 (10 verbs matched), and 320 (10) with predicative form; it appears in KKS 882 (8) with attributive form; terms in boxes are the 7 elements; $\alpha_1 = \{ru, ta, iku\}$; $\alpha_2 = \{aru, iku, kuru, shinau, \dots\}$; $\alpha_3 = \{rareru, nai, \dots\}$; $\beta_1 = \{u, darō, dearō, deshō, \dots\}$; $\beta_2 = \{ka, na, yo, zo, sa, \dots\}$; $LK = \{te, de, ni, \dots\}$; and $DN = \{no, koto\}$.

5 Conclusion

The purpose of this paper is to develop a dataset comprising elements in predicate verb phrases of classical Japanese poems and their modern translations, and to verify the differences between the verb phrases in OP and CT. We sought to classify all elements in CT into syntactic, contextual, and other elements. To simplify the analysis, we focused only on poems that contained verbs in plain form. We found that most elements in CT predicates are based on the literal elements present in the OP. Predicate verb phrases in classical Japanese poetry exhibit various patterns beyond the plain form, necessitating the development of methods to address other complex forms within the current dataset.

Bibliography

- Katagiri, Yoichi (1990) *Gosenwakashu*, Shin Nihon koten bungaku taikai, Tokyo: Iwanami Shoten. Katano, Tatsuro and Yoichi Matsuno (1993) *Senzaiwakashu*, Shin Nihon koten bungaku taikai, Tokyo: Iwanami Shoten.
- Kato, Yasuhide (1986) “Bunmatsu ni shiyō sareru dōshi no imi yōhō no shidō (Teaching methods of meanings and uses of verbs used as a predicate)”, *Nihongogaku*, Vol. 5, No. 4, pp. 55–63.
- Kawamura, Teruo, Yoshio Kashiwagi, and Shigenori Kudo (1989) *Kinyōwakashu*, Shikawakashu, Shin Nihon koten bungaku taikai, Tokyo: Iwanami Shoten.
- Kojima, Noriyuki and Eizō Arai (1989) *Kokinwakashu*, Vol. 5 of Shin-Nihon bungaku taikai (A new collection of Japanese literature), Tokyo: Iwanami shoten.

Komachiya, Teruhiko (1990) *Shuiwakashu*, Shin Nihon koten bungaku taikai, Tokyo: Iwanami Shoten. Kubota, Jun and Yoshinobu Hirata (1994) *Goshu iwakashu*, Shin Nihon koten bungaku taikai, Tokyo: Iwanami Shoten.

Kubota, Jun (1979) *Shinkokinwakashu*, Shincho Nihon Koten Shu-sei, Tokyo: Shinchosha.

Nakamura, Yasuo, Yoshihiko Tachikawa, and Mayuko Sugita (1999) *Kokubungaku kenkyu-shiryokan detabesu koten korekushon "Niju ichidaishu" Shohobanbon CD-ROM* (Database Collection by National

Institute of Japanese Literature "Niju ichidaishu" the Shoho edition CD-ROM): Iwanami Shoten. Suzuki, Shigeyuki (1965) "Gendai nihongo no dōshi no tensu —iikiri no jutsugo ni tsukawareta baai— (Tense of verbs of Modern Japanese: the case of verbs at the end of the predicate.", in *Kotoba no kenkyu-2* (Study of language), Vol. 2 of Report of the National Language Research Institute, Tokyo: Shuei shuppan, pp. 1–38.

Takahashi, Taro (1983) "Suru tomo shita tomo ieru toki (The cases that can be expressed by both 'suru' and 'shita'", in *Kindaichi Haruhiko Hakushi koki kinen ronbunshu*, Vol. 2, Tokyo: Sanseido, pp. 405–34.

Yamamoto, Hirofumi and Bor Hodo'scek (2021) "Hachidaishu vocabulary dataset", <https://doi.org/10.5281/zenodo.4744170>.

Yamamoto, Hirofumi Hilo (2005) "A Mathematical Analysis of the Connotations of Classical Japanese Poetic Vocabulary", Ph.D. dissertation, Australian National University.

Panels

Digital Resources in Buddhist Studies in Taiwan – a Progress Report

Hsiang, Jieh

jieh.hsiang@gmail.com
National Taiwan University, Taiwan

Hung, Jen-Jou

jenjou.hung@dila.edu.tw
Dharma Drum Institute of Liberal Arts

Hung, I-Mei

yimay0519@gmail.com
National Taiwan University, Taiwan

Ting, Pei-Feng

jerryting@ntu.edu.tw
National Taiwan University, Taiwan

Lo, Hao-Cheng

austenpsy@gmail.com
National Taiwan University, Taiwan

Digital Resources of Buddhist Studies in Taiwan – a Progress Report

Moderator: Jieh Hsiang

Abstract

This panel comprises three presentations, centered around some of the current research activities in Taiwan concerning digital resources of Buddhist Studies. In the first talk we present NTU Digital Library of Buddhist Studies (DLBS). While there are a number of online databases of Sutra texts (such as CBETA and SAT), DLBS focuses on bibliographic records of books and journal articles and is among the most accessed online resources of Buddhist Studies. In this talk, we should focus on recent developments such as a 5-language authority database and a database of authors. Also discussed will be our work linking journal citations of Scripture verses with their original texts in CBETA.

Since the release of the new DLBS website in 2006, it has attracted more than 10,000 daily users and more than 40,000,000 of total user sessions from 239 countries. In the second talk, we describe a study of user behavior analysis. We present the methods we designed and use them to analyze the user log of 6 months from June 2022 to November 2022. We shall present the methodologies and some of the discoveries. A study of the entire user log will

reveal many interesting findings of user behavior in DLBS such as the change in research trends over time, differences of sutra usages and research focus among regions, etc.

The third talk showcases our recent work on a context analysis system that incorporates biographic data into Buddhist Temple Gazetteers (BTG). DILA has recently transcribed the complete set of BTG and made it available to the public. However, the sheer volume has made it difficult to use collectively. By using DocuSky to build the context analysis system, we can discover and explore the multiple contexts implicitly embedded in the gazetteers and the relationships among temples and prosopography.

Jieh Hsiang

Department of Computer Science,
Research Center for Digital Humanities
National Taiwan University

Pei-Feng Ting

The DLBS Project, NTU Library
National Taiwan University

The Digital Library of Buddhist Studies (DLBS, <https://buddhism.lib.ntu.edu.tw>) at National Taiwan University (NTU) aims to provide readers worldwide with a rich collection of literature and materials regarding Buddhist Studies. DLBS was founded in 1995 by Venerable Shih Heng-Ching, Professor Emeritus of Philosophy at NTU. It is currently (since 2005) under the supervision of Jieh Hsiang, Distinguished Professor of Computer Science and Director of the Research Center for Digital Humanities of NTU. While many Buddhism-related databases emphasize full-text transcription and tagging of Buddhist Scriptures, DLBS focuses on research material such as journal papers and books on Buddhism. Thus, DLBS is an ideal companion for scholars who use the Sutra databases for research. DLBS currently has more than 470,000 bibliographic records of books and journal articles, among which more than 117,000 have downloadable full-text files. Through linking with other online databases, it also provides Buddhist Scriptures in different languages (images and transcribed full-texts) as well as language teaching tools in Tibetan, Sanskrit, and Pali. The journal articles are collected from 9,595 journals, and we track journal databases and websites regularly to keep our bibliographic records updated.

Although a significant portion of materials in DLBS are in Chinese, Japanese, and English, it also contains bib records in 44 other languages. In addition to keyword search and post-classification of search results, DLBS also features a five-language (Chinese, English, Japanese, German, French) thesaurus of Buddhist terms and an Author Authority file that contains more than 110,000 authors.

The DLBS website has interfaces in Chinese, English, and Japanese. Since its inauguration in 2006, it has attracted more than 40,000,000 visits/sessions from 239 countries and has more than 10,000 visitors daily. It is among the most visited websites on Buddhist Studies.

A feature recently added is a cross-citation mechanism between journal articles in DLBS and the sutra database CBETA. We identify the quotation of verses and the scripture that they appeared and link the citations back to CBETA. CBETA can also query DLBS via API.

In this talk, we shall give an overview of the main features of DLBS: its contents and functions. We shall also describe the interaction features between DLBS and CBETA and future work.

Hao-Cheng Lo

Department of Psychology

National Taiwan University

Jieh Hsiang

Department of Computer Science

Research Center for Digital Humanities

National Taiwan University

Understanding user behavior patterns becomes crucial as digital libraries and institutions expand their online presence. The NTU Digital Library of Buddhist Studies (DLBS), accessed by more than 10,000 users daily, has meticulously kept its user records since its inauguration in 2006. This talk describes a study that employs data-driven psycho-informatic methods, machine learning, and statistical techniques, including natural language processing and survival analysis. Our multifaceted approach provides valuable insights into how users interact with the open-access Buddhist book texts, which can enhance the design and functionality of the online library and ultimately improve user experience, contributing to the growing body of knowledge in digital humanities.

As a preliminary attempt made to DLBS, we analyze an extensive dataset comprising eight million browsing logs spanning over a 6 month (June to Nov 2022) period, detailing users' interactions with the bibliographic in DLBS, namely bibliographic information (book identifier, book title, media type, book language, page language, and availability of the full text), IP address information (geolocation), user behavior (visit counts, visit span, and last visit). Our methodological approach integrates descriptive visualizations, text analysis (topic modeling and semantic embedding), latent analysis (NMF and t-SNE), and survival analysis. Visualization techniques, including plots, maps, networks, and word clouds, are employed to facilitate data interpretation. Prior to this analysis, we conducted data processing and cleaning to filter meaningful data. Additionally, to further enable detailed exploration, we have developed an interactive application for filtering the dataset and visualizing outcomes and implementing a recommendation system based on our findings.

Regarding user engagement, we analyze bibliographic information and geolocation data to identify patterns and trends through descriptive visualizations, highlighting areas that may benefit from further improvement or optimization. In examining content popularity, we uncover the most

frequently accessed books and subjects within the library, offering valuable information for digital library curators to focus their efforts on the most appealing content to users.

We investigate the influence of factors such as geographic location, language, and media type preferences on user engagement, which allows us to provide targeted insights into the needs and preferences of different user segments. This understanding can aid digital library designers and administrators in tailoring their offerings to better serve their user base. Through applying NMF and t-SNE techniques, we uncover latent relationships between users and books on user interactions, identifying user clusters that share similar preferences and behaviors and book clusters that group books commonly accessed by the same user segments.

By performing topic modeling (LDA) on book titles, we reveal prevalent themes and subjects in online Buddhist literature, allowing for a deeper understanding of user interests. We also create word clouds to visually represent the most common keywords and themes within the library's content, providing an easily understandable representation of user preferences.

Our survival analysis uncovers user retention patterns by Kaplan-Meier estimator and Cox proportional hazards models, highlighting factors that contribute to users continuing to engage with the digital library over time. We examine the factors influencing user retention patterns, such as media type, language, and topic. Additionally, we explore whether user retention varies across geographic locations.

Our study offers a threefold contribution to the field of digital humanities. First, our comprehensive methodological approach, which integrates data-driven techniques, demonstrates the versatility and potential of digital tools in advancing humanities research, particularly in the realm of user engagement and retention. Notably, while our research uses the DLBS as a case study, we stress that the analytic techniques employed are not exclusive to this library. Given the universality of these techniques, they can certainly be applied to user studies of any bibliographic database with a well-maintained user record, especially one with a focused research domain (so that the keywords are not too diverse). Second, we have developed an interactive application and recommendation system that improves user experience and satisfaction and highlights the role of digital humanities in enhancing the accessibility and functionality of online libraries. Lastly, through the construction and in-depth analysis of a rich dataset, we provide a strong foundation for future studies in the field while revealing patterns, trends, and factors influencing user engagement and retention.

Our research aligns with diverse practices and interdisciplinary aspects of digital humanities, showcasing the applicability and relevance of data-driven research within this domain. By addressing these critical aspects, our study significantly contributes to the growth and evolution

of data-driven research in the digital humanities, paving the way for further innovations and collaborations in the field.

Jen-Jou Hung

Department of Buddhist Studies

Dharma Drum Institute of Liberal Arts

I-Mei Hung

Research Center for Digital Humanities

National Taiwan University

In Buddhism, Buddha, Dharma, and Sangha are known as the Three Jewels, which can be seen as the three most important elements of Buddhism. As a result, traditional Buddhist literature has always placed a high value on the teachings of Buddha and the activities of the Sangha, resulting in many important works being created. However, the propagation of the Dharma and the translation of scriptures have all taken place in Buddhist temples, yet records centered on these temples have not received much attention, and there are few related works. It was not until the Ming and Qing dynasties that the trend of compiling historical records for Buddhist temples began to emerge in various regions of China. These records, known as "Buddhist Temple Gazetteers" (佛寺志, fo-si-zhi) by modern scholars, cover the history of temples, geographic landscapes, the teachings and actions of masters, temple properties, ancient architecture, important documents, and artistic and cultural creations, forming a unique category of Buddhist literature. Recently, these Gazetteers have been compiled into two book collections, including the 中國佛寺史志彙刊 (*Zhongguo Fosi Shizhi Huikan*) and the 中國佛寺志叢刊 (*Zhongguo fosizhi congkan*). The publication of these two series has rekindled academic interest in the contents of Buddhist Temple Gazetteers, which have become a key resource for studying the history of Buddhist temples during the Ming and Qing dynasties.

Since 2008, the Digital Archives Team of Dharma Drum Institute of Liberal Arts (DILA) has been working on the full digitization of two collections. They have made the digitization of Buddhist Temple Gazetteers publicly available on the internet for free use. The content of this digital resource is arranged according to the original book's order without being categorized by specific topics. Therefore, when using the digital text of the Buddhist Temple Gazetteers, readers often encounter difficulties in applying and analyzing the information systematically due to the problems of complexity and inconsistency in the structure of the content. Therefore, we propose to improve this problem by establishing a "thematic context analysis system." This study will use biographical data to construct this system. There are two reasons for this: first, after consulting with experts and Buddhist scholars, we found that the biographical data contained in the Buddhist Temple Gazetteers has significant value for Buddhist studies. Second, based on our preliminary statistics, biographical

data in the Gazetteers accounts for approximately 17% of the overall content, making it the second-largest category after artistic and cultural data.

In terms of technology for constructing the system, we chose the DocuSky platform developed by the Digital Humanities Research Center at National Taiwan University as a template to create a highly exploratory thematic context analysis system. By linking cross datasets (biographical data of individuals and Buddhist temple records), the system allows the biographical data of individuals to not only contain the information of the individual subjects but also preserves the meaningful connections between individuals and the associated temple records. This also enables readers to conduct in-depth exploration and research of Buddhist figures through the time, space, and content structure of Temple Gazetteers. By utilizing the functions provided by the system, such as post-classification, data reference, and visualization, readers can engage in interactive analysis of multiple contexts. This approach offers a new research perspective and methodology for the study of Buddhist figures. In this presentation, we will explain in detail the development and subsequent application of the "Buddhist Temple Gazetteers Biographical Data Context Analysis System".

The DocuSky platform which this system chose as the development template, adopts a set of structural language used to explore the application boundary of digital humanities. The template feature focuses on describing the document structure and the application method of the collective characteristics of the documents and incorporates agile experimental tests, etc. For practical considerations, we believe that this process and model based on deconstructing a comprehensive full-text database to establish the context analysis system can serve as a reference for similar DH projects.

Possibilities of Digital Social Science and Data-Driven Studies

Shao, Hsuan-Lei

hlshao2@gmail.com

National Taiwan Normal University, Taiwan

Huang, Sieh-Chien

schhuang@ntu.edu.tw

National Taiwan University, Taiwan

Chao, Shiau-Fang

sfchao@ntu.edu.tw

National Taiwan University, Taiwan

Yeh, Yu-Chun

eagleuu6@gmail.com

National Taiwan Normal University, Taiwan

Wu, Chia-Chia

igu19940613@gmail.com

National Taiwan Normal University, Taiwan

Chang, Gia-Ming

rogergo9929@gmail.com

National Taiwan Normal University, Taiwan

Abstract

The digital revolution has brought about a wealth of possibilities for social scientists and humanities scholars to explore and analyze data in innovative ways. With the advent of new technologies and digital tools, there has been an increased interest in data-driven studies, which have the potential to transform our understanding of society, mass communication, politics, and human behavior.

This panel is organized by the "Center of China Studies, NTNU (main project: The Knowledge Database/ Graph of China-studies, <https://ntnu2021.herokuapp.com/>) and "Center of Digital Legal Studies, NTU". We will bring together three papers from the member of the centers, from fields of digital social science, digital gerontology, and digital politics--all of them are based on digital humanities to explore the possibilities of information science and data-driven studies for advancing digital humanities research. Although there papers are from different topics and fields, the panel will address questions such as: What are the key opportunities and challenges of using data-driven approaches in digital humanities (and other fields) research? How can digital social science methods be used to analyze large-scale data sets to gain insights into social and cultural phenomena? What are the ethical considerations involved in using data-driven approaches in humanities research?

Through their presentations and discussions, the panel will highlight the importance of collaboration and interdisciplinary approaches for advancing digital humanities research in the era of big data. One area where digital social science and data-driven studies have made a particularly significant impact is in the field of digital humanities. Digital humanities is an interdisciplinary field that combines traditional humanities disciplines with digital technologies, and it has become increasingly popular in recent years as scholars have recognized the potential of digital tools and techniques for advancing humanities research. Digital humanities can encompass a wide range of

topics and methods, from data visualization and text mining to digital archiving and online publishing.

The panel on "Possibilities of Digital Social Science and Data-Driven Studies" will explore the ways in which digital technologies are transforming social science research, and the ways in which data-driven studies are contributing to our understanding of the social world. We will also examine the potential of digital humanities as a means of advancing humanities research, and the ways in which digital tools and techniques can be used to support interdisciplinary research and collaboration.

Topics that may be addressed in this panel could include:

1. The use of digital technologies in social science research, including data mining, text analysis, and NLP skills.

2. The challenges and opportunities of working with large datasets, including issues related to data collecting, data quality, and data management.

3. The role of digital humanities in advancing humanities research, including the use of sociology, mass communication and Politics.

Overall, this panel will provide an opportunity for scholars to share their research and ideas on the possibilities of digital social science and data-driven studies, and the ways in which they are transforming the social sciences and humanities.

Keywords: digital social science, information technology, text mining, digital gerontology, Data-Driven Studies

Panel Papers

The panel is including three papers:

Aging Society and Digital Gerontology: Applying Machine Learning to Analyzing Taiwanese Public Attitudes of Elderly Healthcare Spending

Wu, Chia-Chia(1). Chao, Shiau-Fang (2), Sieh-Chuen (2)¹

Organization(s):

1: National Taiwan Normal University, Taiwan;

2: National Taiwan University, Taiwan

A. Research Background:

An aging society is a global trend. According to the National Development Council, Taiwan became an aging society in 1993. Since the implementation of the National

Health Insurance in March 1995, medical expenses have increased significantly. According to the data published by the National Health Insurance Administration under the Ministry of Health and Welfare in 2018, the elderly aged 65 and above accounted for 32.9% of the total health insurance expenditures, and their hospitalization days and total medical expenses accounted for 45.8% and 48.1%, respectively. The "Achievements and Challenges of the National Health Insurance System" research report commissioned by the Taiwan Economic Association in 2015 pointed out that the hospitalization rate, emergency department utilization rate, and medication use rate of the elderly aged 65 and above were higher than those of other age groups, and their medical expenses were also much higher than those of other age groups.

There are also dissenting voices in society regarding the "medical expenditure situation of the elderly population." Researchers observed that in late 2019, during the COVID-19 epidemic, the Taiwanese public forum "PTT" frequently discussed whether "the elderly occupy too much medical resources." This study aims to conduct sentiment analysis of the PTT gossip forum corpus from 2019 to 2022 during the epidemic period using machine learning and manual annotation. The purpose is to understand what kind of ideology is presented in the current Taiwanese public opinion towards the elderly's use of the majority of medical resources, whether there is a positive or negative aspect to it, and how to respond to such social phenomena. Through this research, the study aims to understand the Taiwanese public's attitude towards the elderly's medical expenses and hopes to find opportunities to mitigate intergenerational conflict.

B. Research Questions:

1 What are the key terms and their frequency used in discussing "the use of medical resources by elderly people" on the PTT forum?

2 What is the main content discussed on the PTT forum regarding "the use of medical resources by elderly people"?

3 What kind of ideology is presented when discussing "the use of medical resources by elderly people" on the PTT forum through sentiment analysis?

4 What attitudes are presented when discussing "the use of medical resources by elderly people" on the PTT forum through sentiment analysis?

C. Research Method:

The study employs natural language processing (NLP) methods to conduct sentiment analysis on textual data. Since the COVID-19 pandemic began in late 2019 and

is still prevalent globally in April 2023, the study uses the corpus collected from the PTT gossip forum between December 1, 2019, and December 1, 2022, for sentiment analysis. The study applies the bag-of-words method to extract keywords from the textual data, which are then transformed into feature vectors that can be processed using machine learning techniques such as neural networks. The trained model uses the feature vectors as input and their corresponding sentiment labels as output. The trained model is then applied to new textual data for sentiment analysis. The experimental results of the sentiment analysis, including accuracy, precision, recall, and F1 score of the text classification, are presented.

D. Predicted Contribution

This research suggests that natural language processing (NLP) techniques can be used to extract insights from large volumes of textual data, such as online discussions on internet forum PTT. These insights can help researchers in the humanities to better understand and analyze social phenomena, cultural trends, and human behavior.

Japan and Taiwan are both experiencing aging societies, with their populations aging at a rapid pace. In Japan, the aging population has been a pressing issue for several decades, and the government has implemented various policies to address the challenges posed by an aging society, such as promoting elderly employment and encouraging immigration. Similarly, in Taiwan, the proportion of elderly citizens has increased steadily in recent years, and the government has also implemented policies aimed at promoting healthy aging, such as increasing healthcare resources and providing subsidies for long-term care.

As both Japan and Taiwan continue to grapple with the challenges of aging populations, there is a growing interest in utilizing data-driven approaches to better understand the needs and behaviors of elderly populations. Digital humanities and related fields have the potential to play a significant role in this endeavor, providing researchers with powerful tools for analyzing large amounts of data and gaining new insights into the experiences of aging individuals. By combining data-driven approaches with insights from the humanities and social sciences, researchers can gain a more holistic understanding of the complex social and cultural factors that shape the experiences of elderly individuals in these societies.

This relation of this research with "the possibilities for data-driven humanities", where scholars in the humanities can use digital tools and computational methods to analyze and interpret large amounts of data. This is closely related to the field of digital humanities, which involves the use of digital technologies and computational methods in humanities research. By combining the knowledge and

methods of the humanities with those of computer science, digital humanities can offer new ways of approaching both aging society and social media, which has the potential to unlock new insights and understanding, and to expand the possibilities for data-driven humanities.

Rus-Ukrainian War on Chinese Weibo: Topic Modeling and Digital Politics

Yeh, Yu-Chun ; Shao, Hsuan-Lei²
National Taiwan Normal University, Taiwan;

A. Research Background

The Russia-Ukraine war started on February 16, 2022. At first, Russia called it a "special military operation", until the bilateral officially declared war on February 24, 2022, and it continues to this day. Since the outbreak of the war six months ago, the West has made it clear in its diplomatic, military and economic stances that it will provide support to Ukraine and call on the international community to sanction Russia. On the contrary, China, as Russia's traditional ally in international politics, has an unclear position at the beginning of the war, and does not forcefully express its anti-American hegemony and anti-Western attitude until the middle and late stages of the war. In the early stage, the China officials keep calm and act like the peacemaker, while in the later, they tend to speak out for Russia. In addition, China's military actions in the Pacific have clearly demonstrated its status as a military ally with Russia, increasing the insecurity in the Asia-Pacific region.

B. Research Method

Based on the position and timing changes of the Chinese Communist Party media on the Russian-Ukrainian war, we analyze 30,475 Weibo posts related to the war between February 21 and June 19, 2022, using three approaches: machine learning, topic analysis, and timing analysis. The reason why choosing Weibo as study data is that Chinese people have restrictions on free speech and Internet use, so they mainly use Weibo as a social networking platform, at the same time, Weibo is also regarded as the Chinese version of Facebook. In this study, we classify Weibo's posts into five categories: European diplomacy, China's stance, warfare, UK-US Diplomacy, and economic impact, examine the posting and interaction situation between official and personal accounts, and compare the data with the timeline of the conflict.

C. Research Content

The research reveals that posts related tragedy of the Russia- Ukraine war tend to attract more comments and sharing, and express sadness about the cruelty of the war. Furthermore, posts originating from official accounts tend to spread more widely, as they are often sharing by other accounts. However, This happens less often with private accounts. In conclusion, the study identifies five results. First, the Russia-Ukraine war may become protracted, while Putin's political reputation suffers a rapid decline. Second, the internet has strengthened Ukraine's national identity. Third, Russia may increase cooperation with China, which could pose a threat to stability in the Asia-Pacific region. Fourth, China may learn from the Russia-Ukraine war and prepare for resource competition in the Asia-Pacific region. Fifth, the internet and media have become crucial resources in wartime, and even official and public web traffic can influence a country's assessment of the situation. Therefore, how to effectively use and manage online resources, and how to understand public trends to prevent conflicts, must be a top priority for national security and defense in the future.

In light of these findings, it is essential to continue analyzing online data to gain insight into global conflicts and trends. Governments and individuals should work to prevent the spread of misinformation, which can fuel conflicts and undermine efforts to maintain peace. From the war between Russia and Ukraine, the critical role of the internet and social media in shaping the world's perception of conflicts, and underscores the need for responsible online behavior and effective management.

D. Expected Contribution

In conclusion, this study highlights the importance of incorporating digital humanities approaches into research on international relations and geopolitics. By applying text-mining and machine learning to analyze corpora from various sources, researchers can gain valuable insights into the strategic intentions of governments and other actors in the global political landscape. The potential of digital humanities to inform decision-making processes and improve our understanding of complex issues underscores the importance of continued investment in this field.

3. Digital Political Science and Xi Jinping's Leadership Style: A Comparative

Analysis with Mao Zedong and Deng Xiaoping

Chang, Gia-Ming; Shao, Hsuan-Lei³
National Taiwan Normal University, Taiwan;

A. Research Background:

"Chinese Communist Party (CCP) political dynamics" has always been an indispensable topic in modern Chinese studies. Theoretically, CCP politics are typically characterized by a power structure where policies are determined by the will of the top leader, commonly referred to as "democratic centralism" in traditional research. This phenomenon was particularly prominent during the eras of Mao Zedong and Deng Xiaoping. However, starting from the Jiang Zemin and Hu Jintao eras, the power of the top leader was diluted due to political unwritten rules set by Deng Xiaoping before his retirement, such as "collective leadership, term limits, designating successor from the next generation, and separating party and government roles." This led to a trend of institutionalization and power decentralization in the CCP political structure, where decisions were no longer made solely by the top leader, but by a group of 6 to 8 senior officials from the State Council who also held positions as members of the Politburo Standing Committee, the elite of the party, thus known as "collective leadership."

Therefore, some observers even consider this as a prelude to the democratization of the CCP. However, this trend was reversed when Xi Jinping, the fifth-generation top leader of the CCP, passed the "Constitutional Amendment" at the First Session of the 13th National People's Congress in March 2018, removing term limits for the President and Vice President, breaking the paradigm set by Deng Xiaoping after the reform and opening. This led to a further concentration of power in the top leader, and in the same year, the "Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era" was incorporated into the party constitution during the 19th National Congress of the CCP, making Xi the leader who had his own name written in the party constitution, and also the core leader supported by the party, following in the footsteps of Mao Zedong and Deng Xiaoping. Up until the 20th National Congress of the CCP in October 2022, Xi Jinping broke with convention and began his third term.

From these trends, it is evident that Xi Jinping will become the long-term leader of the CCP in the future. Therefore, understanding Xi Jinping's leadership style has become a crucial issue in contemporary Chinese studies. In traditional Chinese studies, research on the character traits and political personalities of CCP leaders has never

ceased. There have even been approaches that delve into the political psychology of Chinese people (Bai Luxun, "The Political Psychology of Chinese People"). However, before the reform and opening in 1979, China was isolated from the world, and Chinese researchers often had to rely on official documents and information provided by specific individuals to assess the political situation of the CCP and the path designated by the leaders. This method is an extension of Kremlinology, a study of the Soviet Union, and often requires long-term investment of time by experts to become "familiar" with the styles of these leaders. As a result, only a few experts can master this approach, and it is also difficult to accumulate knowledge through this method.

This study, however, adopts the research method of "computational politics" and text mining, introducing an information-based approach to analyze CCP politics. It is capable of describing thought processes in a programmatic or quantitative manner, enabling rapid familiarity and knowledge accumulation. Of course, this method is still at the forefront of development and has faced skepticism and doubts from some scholars. It also has its own unresolved problems. Therefore, this paper also incorporates traditional literature review methods and historical analysis methods, using historical resolution and speeches passed by three generations of CCP leaders. (Mao Zedong, Deng Xiaoping, and Xi Jinping)

B. Research Problem:

- 1 What is the different leadership style of three generations of CCP leaders?
- 2 What is meaning of every eras of historical resolution latent topic?
- 3 What is connection between Xi Jinping leadership style and his policy?

C. Research Method:

This research is distinctive in its application of "computational political science" and "text mining" techniques to Chinese studies, combining traditional text analysis with information technology. Specifically, this study utilizes the TF-IDF (Term Frequency - Inverted Document Frequency) algorithm and TextRank algorithm to preprocess the documents, followed by Latent Dirichlet Allocation (LDA) topic modeling to explore latent topics within historical documents associated with different leaders during various periods. This approach aims to identify ideological tendencies and leadership styles of leaders during different eras, while also incorporating traditional literature for further analysis.

D. Predict contribution

This Study would use a “Data-Driven” method, which is based on the official documents of the CCP leaders, to describe their ideology and leadership styles. This method can find new application and possibilities for the “Digital Humanities”, and help us understand their impact on China from a different side.

Furthermore, it seeks to analyze the variations in leadership styles and personalities among the three generations of leaders. The study also attempts to provide a specific analysis and organization of the content of historical resolution, in order to implement a new interdisciplinary approach to Chinese studies.

Bibliography

Paper 2:

Brady, A. M. (2009). *Marketing dictatorship: Propaganda and thought work in contemporary China*. Rowman & Littlefield Publishers.

Creemers, R. (2017). Cyber China: Upgrading propaganda, public opinion work and social management for the twenty-first century. *Journal of contemporary China*, 26(103), 85-100.

Editor, “Ukraine war in maps: Tracking the Russian invasion”, BBC NEWS, <https://www.bbc.com/news/world-europe-60506682>, retrieve date: 2022/9/22

Fedor, J., Lewis, S., & Zhurzhenko, T. (2017). Introduction: War and Memory in Russia, Ukraine, and Belarus. In *War and Memory in Russia, Ukraine and Belarus* (pp. 1-40). Palgrave Macmillan, Cham.

LAGERKVIST*, J. O. H. A. N. (2008). Internet ideotainment in the PRC: National responses to cultural globalization. *Journal of Contemporary China*, 17(54), 121-140.

MacKinnon, R. (2010). *Networked authoritarianism in China and beyond: Implications for global internet freedom. Liberation Technology in Authoritarian Regimes*, Stanford University.

Michelle Fong, ‘China monitors the Internet and the public pays the bill’, *Global Voices*, (29 July 2014), <http://advocacy.globalvoicesonline.org/2014/07/29/china-monitors-the-internet-and-the-public-pays-the-bill/>, retrieve date: 2022/10/01

Paper 3:

Frank, Dikötter, (2016). *The Cultural Revolution: A People's History 1962-1976*.

Frank, Dikötter, (2021). *Mao's Great Famine: The History of China's Most Devastating Catastrophe, 1958-1962*.

Frank, Dikötter, (2018). *The Tragedy of Liberation: A History of the Communist Revolution, 1945-1957*.

Ezra F. Vogel, Belknap Press (2013). *Deng Xiaoping and the Transformation of China*

Kevin Rudd. *The Avoidable War: The Dangers of a Catastrophic Conflict between the US and Xi Jinping's China*.

Notes

1. Corresponding author: schhuang@ntu.edu.tw
2. Corresponding author: hlshao2@gmail.com
3. Corresponding author: hlshao2@gmail.com

From Documents to DocuSky—Practice and Application

Tu, Hsieh-Chang

hsieh.chang@gmail.com
National Taiwan University, Taiwan

Hu, Chi-Jui

huchijui@cc.ncue.edu.tw
National Changhua University of Education, Taiwan

Kuo, Chih-Wen

ziwenkuo@mail.ncyu.edu.tw
National Chiayi University, Taiwan

Huang, Chia-Hung

orange1052110128@gmail.com
National Taiwan University, Taiwan

DocuSky (<https://docusky.org.tw/DocuSky/home/>) is a personalized digital humanities research platform jointly developed by the Digital Humanities Research Center and the Digital Humanities Laboratory of the Department of Computer Science and Information Engineering at National Taiwan University. As a multi-integrated platform, DocuSky not only integrates resources and tools of different fields but also provides maximum flexibility of use to achieve the concept of “openness”, “freedom” and “autonomy” in digital humanities services. The realization of this concept is mainly based on the platform’s open architecture and personalized core service content, so as to promote diverse and rich digital humanities research.

Humanities scholars have their own preferred texts. They often add metadata to each text and annotate the text content for deeper analysis. For instance, suppose each text has a metadata field that describes the text author. One

may want to get the distribution of authors in a specific collection of texts. On the other hand, suppose each text contains annotated tags that indicate the places mentioned in the text. Given any collection of tagged texts, one may want to display all the places on a geo-map and analyze the place names that occur most frequently. DocuSky is a research platform that offers a solution to this problem (DocuSky, n.d.; Tu, 2018).

In DocuSky, a *document* refers to text that includes metadata and annotated tags. *Post-classification* is a technique used to generate the distribution of features in a metadata field or annotated tags within a collection of documents. With DocuSky, users can create personal databases that support text retrieval, post-classification, and data visualization.

Fig. 1 illustrates the process of using DocuSky. A user can use *DocuSky tools* to download texts from online text repositories (mostly in Chinese), add metadata and tags to create documents, and then compile all the documents into a structured XML file in DocuXml format. This file can then be used to build a DocuSky database. Once the database is constructed, various DocuSky tools can be applied to search, analyze, and visualize its content.

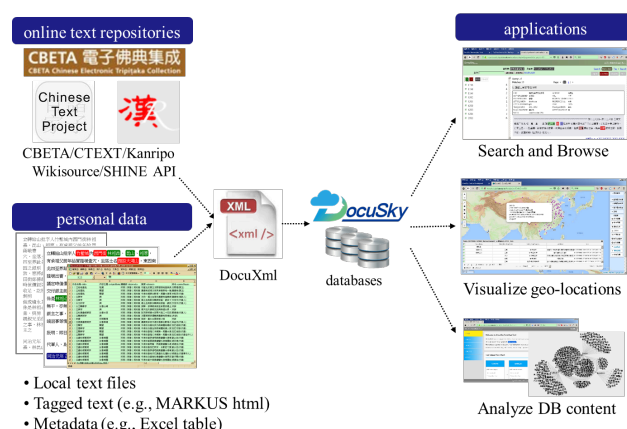


Fig. 1. DocuSky allows one to compose texts, metadata, and tags to build a database for analysis.

The first article of this session is to convert the biographical information of Chinese scholars from the Chinese Biographical Database (CBDB, <https://projects.iq.harvard.edu/cbdb>) to build it into the Chinese Honored-class Scholar (進士 Jinshi) Database. Through the various functions of the DocuSky cloud database, the possibilities of group biography research are expanded. In the second article, the *Mackay Diary* is reorganized from the perspective of humanities research and added with metadata and tags to build a public repository of the Mackay Diary. The third article uses the textual material of the 19th century China medical missionaries, the China Medical

Missionary Journal, and transforms it into a DocuGIS layer to present the footprints of the medical missionaries in China in a textual and cartographic way. The fourth article presents a digital innovation practice that extends from the DocuSky database of “Outline of Astronomy” (天文略) from “Tongzhi” (通志). It is also combined with content from “Butieng” (Song of the Sky Pacers, 步天歌). Through GIS and visualization technology, the interactive application Tongzhi Skymap provides a new approach to research in Chinese astronomy.

DocuSky shares the feature of open architecture. Through cross-country, cross-domain, and cross-resource standard dialogues, it has been able to interact and interface with many large online repositories, such as CBDB, Chinese Text Project (CTEXT, <https://ctext.org/>), Kanseki Repository (Kanripo, <https://www.kanripo.org/>), Chinese Buddhist Electronic Text Association (CBETA, <https://www.cbeta.org/>), Research Infrastructure for the Study of Eurasia (RISE, <https://rise.mpiwg-berlin.mpg.de/>), WikiSource, etc. In addition to different structures of resource transfer, the open architecture built on the standard specification also makes public participation possible. The interoperability approach also breaks through the traditional paradigm of humanities research and allows researchers to browse materials from multiple perspectives. With the help of various digital tools, users can grasp the overall situation from the textual information, interpretation information, and enlightening meaning information through interactive operations. Between distant reading and near reading, and create new digital humanities research topics and viewpoints.

Abstract of Papers

1. Building a CBDB Jinshi database with DocuSky (Hsieh-Chang Tu)

CBDB stands for *China Biographical Database*. It is the product of a project that aims to collect distinguishable historical persons in traditional China. One may regard CBDB as a large-scale *person authority database* that records the name, addresses, offices, kinships, social status as well as social relationships of a person as characteristics. An *authority database* assigns a unique identifier to each collected object and promises that the object identifier won't be changed in the future. This promise and the openness make it possible for third-party applications. Due to the restriction of historical materials, a lot of persons collected by CBDB have little research impact. It is often useful, however, to extract an interesting subset of CBDB for further analysis. For instance, *Jinshi* (進士) is an honored class of scholars who passed the highest level of imperial examination in traditional China. It is interesting to extract all the Jinshis data from CBDB for deeper analysis.

CBDB has another problem for researchers to use. Although it provides several implementations to help

its users find specific persons, it is often difficult to use the systems to get collective properties of persons. *Post-classification* is a technique that classifies a given group of objects (often obtained as a search result) by their properties. It yields the *distributions* of these objects by their *features*. For instance, if all objects in the group have the feature of “publishing year,” taking post-classification over a group of objects yields the distribution of publishing years over the objects. The distribution of objects by their publishing years is simply an ordered list of publishing years where each year is associated with a number to denote the number of objects published in that year. Post-classification can also be helpful to narrow down a search result, say to only retrieve objects with specific publishing years.

DocuSky is a research platform developed by the NTU Research Center for Digital Humanities that allows one to build personal databases that support text retrieval, post-classification, and data visualization (Tu, 2018). By regarding person characteristics as object features, we can convert CBDB data to the DocuXml format (DocuXml 1.3 Scheme: https://hackmd.io/@DocuSky/BksNFnEK_) ready to build a DocuSky database. Once the database is constructed, one can search a desirable group of persons with text retrieval, narrow down the search result, and then apply post-classification to the final result to get feature distributions of the retrieved persons.

In this paper, we discuss how to extract all the Jinshi data from CBDB to build a DocuSky version of CBDB Jinshis. In short, we download and setup the entire database, select all the identifiers which correspond to qualified Jinshis, use CBDB API to download person data, and then convert the data to DocuXml for building the DocuSky version CBDB Jinshi database. This database is open for public access (https://doi.org/10.6681/NTURCDH.DB_DocuSkyCBDBJinshi/Text). We show that, with this newly built database, one can easily get collective properties of Jinshis and present the results in a visualized way. For instance, one may plot Fig. 2 in merely a couple of minutes. This figure integrates visualizations of nine distributions over five collections of Jinshis (each collection corresponds to a search result). It provides in-depth information that is difficult to obtain through traditional methods.

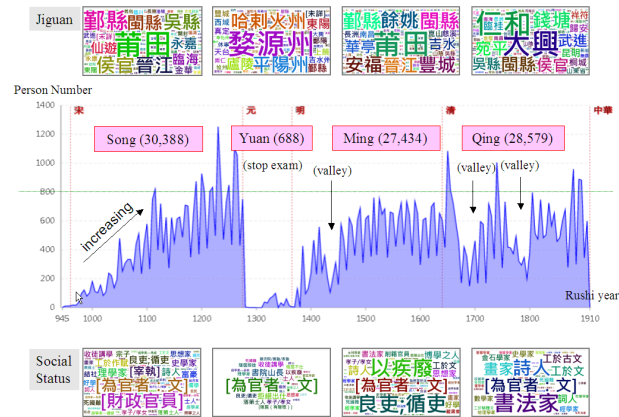


Fig. 2. Feature distribution of Chinese Jinshis from Song to Qing dynasties

The central block of Fig. 2 is obtained by taking post-classification over all the Jinshis in the four dynasties Song (宋, AD960-1279), Yuan (元, AD1279-1368), Ming (明, AD1368-1644), and Qing (清, AD1644-1911). It shows the distribution of the Rushi (入仕, entering government service) year of these persons. CBDB collects 30388, 688, 27434, and 28579 Jinshis in the four dynasties, respectively. It's easy to see that the number of Jinshis overall increases in the Song dynasty. The number of Jinshis drops drastically in the Yuan dynasty due to exam stop. There is a “valley” in the Ming dynasty and two clear valleys in the Qing dynasties. In addition, Fig. 2 adopts *word clouds* to visualize the major Jiguan (籍貫, place of origin or birth) and social status of the Jinshis in each dynasty (4 clouds in the top area and 4 in the bottom area, each shows the major feature items of Jinshis in a dynasty). A *word cloud* displays major feature items (i.e., items with the most numbers) in a distribution such that the item with larger number is presented with a larger font. For instance, the top-left block in Fig. 2 shows the major Jiguans in the distribution yielded from taking post-classification over Jinshis in the Song dynasty by the feature “Jiguan.” From the top-area blocks it's easy to see that Putien (莆田) were the Jiguan that produced the most Jinshis in both Song and Ming dynasties. On the other hand, Wuyuan County (婺源) and Daxing County (大興) were the most productive Jiguans in the Yuan and Qing dynasties, respectively. The bottom-area blocks show the social status of these Jinshis, where one could have more than one social status. Not surprisingly, most Jinshis worked as civil servants ([為官者: 文]). Many Jinshis were finance officer ([財政官員]) in the Song dynasty. Many were good or efficient official (良吏; 循吏) and numerous deposed on account of illness (以疾廢) in the Ming dynasty. In the Qing dynasty, many Jinshis were calligraphers (書法家) and numerous were poets (詩人).

2. Building A Digital Humanities Database by Yourself: A case Study of Mackay's Diary (Chi-Jui Hu)

The *Diary of Mackay* was written by the Rev. George Leslie Mackay (1844-1901), who was a missionary in Taiwan during the late 19th century. The diary began on Nov. 1st, 1871 when Mackay left San Francisco, and ended on Feb. 12th, 1901. Except the diary of 1883, his diary was digitized by the Presbyterian Church in Taiwan and Aletheia University. In 2002, Aletheia University built a database with the image and full-text database of Mackay's diary. After that, in 2019 the Institute of Taiwan History at Academia Sinica rebuilt another Mackay's Diary database in their Taiwanese Diaries Database. In addition to these databases, a new Chinese version of the *Diary of Mackay* was also published in 2012. With these foundations, we use the DocuSky Collaboration Platform to reorganize the full text of the Chinese version diary as an example and add tags and metadata to rebuild a database with the thought of digital humanities by the DocuSky. In the database from the DocuSky model, we can not only use the digital humanities tools from the DocuSky but also explore the relationship of people that Mackay mentioned in his diary by the tags of PersonName and retrieve the locations that he preached the *Gospel* in Taiwan by the GIS map. Through this research, we hope not only to present Mackay's daily life in his diary but also show how to help researchers, by using the digital humanities platform, rebuild the material they already have and add more useful elementary to their research needs by themselves.

3. Visualizing the Footprints of Medical Missionaries in 19th Century China with DocuGIS (Chih-Wen Kuo)

The *China Medical Missionary Journal* is a journal published by the Medical Missionary Association, which was founded in 1886 and served as a society of medical missionaries working in China. This journal contains a variety of information, such as medical care, environment, and hygiene in China. One of its objectives was to establish a platform for information sharing to assist China medical missionaries in obtaining relevant medical information. In June 1887, the *China Medical Missionary Journal* published an article, "Medical Missionaries to the Chinese," describing the service of China medical missionaries in the 19th century. It detailed the China medical missionaries from 1820 to 1886, including information on their paths to China and the places where they practiced medicine. DocuGIS is a tool for DocuSky to integrate geographic information system, which can link the tags of location names in the text with geospatial information to visualize the geographic information content of the text. Users can operate DocuGIS in two ways. The first way is to use the GeoPort tool to import files with location name information or spatial coordinate information from DocuSky database into DocuGIS. The second way is to set the metadata to

tag the location name information in the text and import it into DocuGIS. In this paper, the location names mentioned in the article "Chinese Missionaries to the Chinese" are recorded with self-defined metadata. Through the DocuGIS visualization of textual geographic information, the trajectory of medical missionaries in China between 1820 and 1886 can be presented in textual maps.

4. Tongzhi Skymap: An Interactive Tool for Exploring Butienge in Chinese Astronomy (Chia-Hung Huang)

Web Application Link: <https://tongzhi-skymap-new.vercel.app/>

DocuSky Database: <https://reurl.cc/eDzVx7>

Throughout history, only two distinct schemes for mapping the stars have enjoyed widespread usage: the Babylonian-Greek one and the Chinese one. Unlike its Western counterpart, Chinese astronomy lacks a geometric model to consistently describe the general appearance of the world (F. R. Stephenson, 2011). As Zheng Qiao, the author of *Tongzhi* (Qiao Zheng, 1161), noted in his book, the charts of Chinese astronomy were often significantly flawed. Due to the limitations of printing technology, written language proved to be the most effective means of transmitting astronomical knowledge during that era, compared to the use of charts. This is why the Butienge (Song of the Sky Pacers, 步天歌) plays an important role in Chinese astronomy. Butienge was an early Chinese star catalogue that was most widely used by folks from the Sui dynasty to the Qing dynasty. Its poetry form allowed people to memorize and locate all of the stars by simply singing it. Based on the Butienge, Zheng Qiao used the song as a mnemonic to enumerate all of the stars and later wrote the *Outline of Astronomy* (天文略) in *Tongzhi* (Comprehensive Treatises, 通志). It was not until Zeng wrote this book that the Butienge was recorded in official documents and preserved until today.

Thanks to the use of Western astronomy and GIS visualization techniques, Chinese astronomy now has access to geometric models, allowing for greater accuracy. Although there are various applications available around the internet, none of them can simultaneously demonstrate the true appearance of the Chinese stars with the Butienge in an interactive way. This highlights the pressing need for the application of informatics techniques to the field of Chinese astronomy.

This interactive web application, *Tongzhi Skymap*, showcases the Chinese stars, the Chinese constellations, the Milky Way, and the contents of Butienge at the same time. To create the *Tongzhi Skymap*, I began by extracting all the Chinese star names from the *Outline of Astronomy* of *Tongzhi*. Then, I obtained the proper names and the celestial coordinates (equatorial coordinates) of each star from *Yi Xiang Kao Cheng* (Ignaz Kögler et al., 1752), written by the Jesuit astronomer-missionary Ignaz Kögler,

Antoine Gaubil, and others in the Qing dynasty. Next, I converted these equatorial coordinates into geographic coordinates to facilitate the plotting of stars on the sphere. Moreover, the star color and the lines in the Chinese constellations were collected from Research into Butieng (Xiao-Lu Zhou, 2004). Lastly, the Milky Way, which was not originally included in the Butieng but was later added to the Outline of Astronomy by Zheng, was also incorporated into the skymap. The skymap was built with the d3.js library (D3js.org, n.d.), which combines powerful visualization components and a data-driven approach to DOM manipulation. To visualize geographic data, I used the d3.js package d3-geo (D3js.org, n.d.) and its extension d3-geo-projection (D3js.org, n.d.).

The development of the Tongzhi Skymap has paved the way for integrating Chinese astronomy with informatics visualization tools. The data from the Outline of Astronomy of Tongzhi has driven the creation of the skymap, which provides users with both macro and micro perspectives when exploring Butieng. By providing a novel approach, Tongzhi Skymap has made a valuable contribution to the field, and finally, has opened up new avenues for further studies on Chinese Astronomy.

Bibliography

Tu, Hsieh-Chang (2018). DocuSky: A Platform for Constructing and Analyzing Personal Text Databases. *Journal of Digital Archives and Digital Humanities*, **2**:71-90.

Author Index

Adachi, Junji	27	Ohta, Shoki	57
Aliakbari, Farzaneh	19	Okuyama, Ryogo	58, 60
Aoyama, Mitsuki	57	Roth, Martin	19
Aubert-Bédouchaud, Julien, Maxime	45	Saito, Yuni	58
Camilleri, Gabriele	47	Sato, Eiichi	58
Cao, Fanghui	12	Shao, Hsuan-Lei	71
Carlino, Salvatore	33	Shao, Hsuan-lei	24
Chang, Gia-Ming	72	Takagi, Miu Nicole	62
Chao, Shiau-Fang	71	Tamborrino, Rosa	19
Chen, Xudong	49, 65	Ting, Pei-Feng	69
Delanaux, Remy	19	Tomita, Masaki	58
Eck, Sebastian Oliver	20	Tu, Hsieh-Chang	76
Fang, Wan-Zhen	23	Wang, Yu-Chun	14
Fukumoto, Takaki	52, 57	Wu, Chia-Chia	72
Hodosawa, Tomowa	58	Yamamoto, Hilofumi	49, 64
Hodošček, Bor	49, 65	Yeh, Yu-Chun	72
Hsiang, Jieh	69		
Hsiao, Yi-chen	24		
Hu, Chi-Jui	76		
Huang, Chia-Hung	76		
Huang, Ling-Yi	23		
Huang, Shu-Ling	13		
Huang, Sieh-Chien	71		
Hung, I-Mei	69		
Hung, I-mei	30		
Hung, Jen-Jou	14, 69		
Kanazashi, Tomoya	54, 58		
Kawase, Akihiro	27		
Kikuchi, Nobuhiko	55		
Kinami, Chieri	27		
Kitamoto, Asanobu	35, 39, 45		
Kuo, Chih-Wen	76		
Lai, Yik Po	15		
Landau, Victoria Gioia Désirée	28		
Lin, Nung-yao	29		
Lín, Shu-Hui	30		
Liu, Chao-Lin	30		
Lo, Hao-Cheng	69		
Lui, Pun Ho	15		
Mazanec, Thomas J.	30		
Minadakis, Nikos	38		
Miyagawa, So	33		
Morozov, Mykola	35		
Moysaki, Georgia	38		
Murai, Hajime	52, 54, 57, 60		
Nagasaki, Kiyonori	39		
Ogawa, Jun	39		
Ohba, Arisa	57		
Ohman, Emily	42, 62		
Ohmukai, Ikki	39		