NLP4DH and IWCLUL 2023

The Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages

Proceedings of the Conference

December 1-3, 2023

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 979-8-89176-012-7

Preface

Textual sources are essential for research in digital humanities. Especially when larger datasets are analyzed, the use of natural language processing (We are delighted to welcome participants to the unique and pioneering hybrid event that combines the International Conference on Natural Language Processing for Digital Humanities (NLP4DH) and the International Workshop on Computational Linguistics for Uralic Languages (IWCLUL). This year marks a significant milestone as we bring together two vibrant communities under one umbrella, fostering an interdisciplinary dialogue and collaboration between digital humanities and computational linguistics, with a special focus on Uralic languages and broader linguistic diversity.

The NLP4DH, in its previous edition, highlighted the crucial role of NLP technologies in addressing the specific needs of digital humanists. The focus was on the application of NLP in exploring non-standard languages, dialects, and historical texts, areas that are often overlooked in mainstream NLP research. The event underscored the importance of bridging the gap between the methodological rigor of NLP and the concrete, data-driven inquiries of digital humanities. This year, we continue to emphasize the synergy between these fields, exploring how advanced NLP tools and methods can be fine-tuned and retrained to better serve the nuanced requirements of humanities research.

On the other hand, IWCLUL has been a cornerstone in the study and preservation of Uralic languages, offering insights into traditional language technology resources and modern computational approaches. The previous editions showcased a diverse range of research, from language-specific studies to comparative analyses across the Uralic language family. This year, in conjunction with NLP4DH, IWCLUL aims to extend its reach and impact, exploring how computational linguistics can contribute to the preservation, understanding, and development of Uralic and other minority languages.

The joint event is a testament to our commitment to interdisciplinary research and the recognition of the importance of linguistic diversity in computational studies. We have a rich program that includes high-quality submissions from both NLP4DH and IWCLUL communities. The presentations range from innovative NLP applications in the humanities to cutting-edge computational techniques in Uralic language studies.

We are particularly excited about the potential outcomes of this collaboration. The intersection of digital humanities and computational linguistics, especially in the context of less-researched languages, opens up new avenues for research and application. We anticipate that this event will not only contribute to academic discourse but also pave the way for practical solutions that benefit language communities, researchers, and practitioners alike.

We extend our heartfelt thanks to all contributors, participants, and organizers who have worked tirelessly to make this event a reality. Your enthusiasm and dedication are the driving forces behind this successful collaboration. We look forward to the fruitful discussions, innovative ideas, and new partnerships that will emerge from this unique gathering.

Welcome to the joint NLP4DH and IWCLUL event – a convergence of digital humanities and computational linguistics, celebrating linguistic diversity and interdisciplinary research.

This event is organized in collaboration with SIGUR, ACL Special Interest Group for Uralic Languages.

Organizing Committee

Organizers (NLP4DH)

Mika Hämäläinen, Metropolia University of Applied Sciences
Emily Öhman, Waseda University
Khalid Alnajjar, Rootroo Ltd
So Miyagawa, National Institute for Japanese Language and Linguistics
Yuri Bizzoni, Aarhus University

Organizers (IWCLUL)

Flammie Pirinen, UiT The Arctic University of Norway
Niko Partanen, University of Helsinki
Jack Rueter, University of Helsinki

Program Committee

Reviewers

Aynat Rubinstein, The Hebrew University of Jerusalem
Leo Leppänen, University of Helsinki
Kenichi Iwatsuki, KTTA
Lidia Pivovarova, University of Helsinki
Linda Wiechetek, University of Tromsø
Jouni Tuominen, University of Helsinki
Mikko Kurimo, Aalto University
Balázs Indig, Eötvös Lorand University
Pierre Magistry, Institut National des Langues et Civilisations Orientales
Yoshifumi Kawasaki, The University of Tokyo
Eetu Mäkelä, University of Helsinki
Timofey Arkhangelskiy, Universität Hamburg
Nicolas Gutehrlé, Université de Franche-Comté
Kaisla Kajava, Aalto University
Joshua Wilbur, University of Tartu
Pascale Moreira, School of Communication and Culture
Miikka Silfverberg, University of British Columbia
Francis Tyers, Indiana University
Anna Dmitrieva, University of Helsinki
Somesh Mohapatra, Massachusetts Institute of Technology
Won Ik Cho, Samsung Advanced Institute of Technology
Shuo Zhang, Bose Corp
Pihla Toivanen, University of Helsinki
Antti Kanner, University of Turku
Jeremy Bradley, Universität Vienna
Aatu Liimatta, University of Helsinki
Sijia Ge, University of Colorado Boulder
Michael Rießler, University of Eastern Finland
Irene Russo, Consiglio Nazionale delle Ricerche
Gechuan Zhang, University College Dublin
Maria Antoniak, Allen Institute for Artificial Intelligence
Thomas Schmidt, Universität Regensburg
Juho Pääkkönen, University of Helsinki
Rogier Blokland, Uppsala University
Jenna Kanerva, University of Turku
Katerina Korre, University of Bologna
Mikko Aulamo, University of Helsinki
Mitsunori Ogihara, University of Miami
Miu Takagi, Waseda University
Quan Duong, University of Helsinki
Daniela Teodorescu, University of Alberta
Erkki Mervaala, Finnish Environment Institute
Joachim Scharloth, Waseda University
Dimosthenis Antypas, Cardiff University
Ayana Niwa, Tokyo Institute of Technology
Heiki-Jaan Kaalep, institute of computer science

Shu Okabe, Univ. Paris-Saclay
Dmitry Nikolaev, University of Stuttgart
Ritwik Bose, The Institute for Human & Machine Cognition
Dongqi Pu, Universität des Saarlandes
Aina Garí, Télécom-Paris
Nils Hjortnaes, Indiana University
László Fejes, Hungarian Research Centre for Linguistics
Ligeti-Nagy Noémi, MTA-PPKE
Tulika Bose, Vivoka
Allison Lahnala, Phillips-Universität Marburg
Gabriel Simmons, University of California
Vilja Hulden, University of Colorado at Boulder
Federico Boschetti, CNR-ILC

Table of Contents

Program

Friday, December 1, 2023

09:30 - 10:00	*Opening and Lightning Talks*
10:00 - 10:30	*Comparison on Heterosexual and Homosexual Woman's Lonely Heart Ads in Taiwan: Taking AllTogether and Lesbian Board on PTT Web Forum as Examples*
10:30 - 11:00	*Automated Generation of Multiple-Choice Cloze Questions for Assessing English Vocabulary Using GPT-turbo 3.5*
11:00 - 11:30	*Banning of ChatGPT from Educational Spaces: A Reddit Perspective*
11:30 - 12:00	*Bootstrapping Moksha-Erzya Neural Machine Translation from Rule-Based Apertium*
12:00 - 13:00	*Lunch*
13:00 - 13:30	*A Question of Confidence: Using OCR Technology for Script analysis*
13:30 - 14:00	*An efficient, open machine translation solution for Buddhist Chinese*
14:00 - 14:30	*Translation from Historical to Contemporary Japanese Using Japanese T5*
14:30 - 14:45	*Coffee break*
14:45 - 15:15	*Fly, fly little Comet! Exploring Subtoken-Level Metaphorical Patterns in Finnish and Hungarian Texts. New Results from the FiHuComet Corpus.*
15:15 - 15:45	*Emil.RuleZ! – An exploratory pilot study of handling a real-life longitudinal email archive*
15:45 - 16:15	*Introducing ChatGPT to a researcher's toolkit: An empirical comparison between rule-based and large language model approach in the context of qualitative content analysis of political texts in Finnish*
16:15 - 16:45	*Efficient and reliable utilization of automated data collection applied to news on climate change*

Saturday, December 2, 2023

09:30 - 10:30	*Keynote: Kyo Kageura*
10:30 - 11:00	*The Stylometry of Maoism: Quantifying the Language of Mao Zedong*
11:00 - 11:30	*Explorative study on verbalizing students' skills with NLP/AI-tool in Digital Living Lab at Laurea UAS, Finland*
11:30 - 12:00	*The Great Digital Humanities Disconnect: The Failure of DH Publishing*
12:00 - 13:00	*Lunch*
13:00 - 13:30	*Emotion-based Morality in Tagalog and English Scenarios (EMoTES-3K): A Parallel Corpus for Explaining (Im)morality of Actions*
13:30 - 14:00	*Understanding Gender Stereotypes in Video Game Character Designs: A Case Study of Honor of Kings*
14:00 - 14:30	*Girlbosses, The Red Pill, and the Anomie and Fatale of Gender Online: Analyzing Posts from r/SuicideWatch on Reddit*
14:30 - 14:45	*Coffee break*
14:45 - 15:15	*Revisiting Authorship Attribution of Tirant lo Blanc Using Parts of Speech n-grams*
15:15 - 15:45	*Explicit References to Social Values in Fairy Tales: A Comparison between Three European Cultures*
15:45 - 16:15	*Readability and Complexity: Diachronic Evolution of Literary Language Across 9000 Novels*
16:15 - 16:30	*Coffee break*
16:30 - 17:30	*SIGUR Business Meeting*

Sunday, December 3, 2023

09:30 - 10:00	*Keynote: Maria Antonia*
10:30 - 11:00	*Statistical Measures for Readability Assessment*
11:00 - 11:30	*Study on the Domain Adaption of Korean Speech Act using Daily Conversation Dataset and Petition Corpus*
11:30 - 12:00	*Bridging the Gap: Demonstrating the Applicability of Linguistic Analysis Tools in Digital Musicology*
12:00 - 13:00	*Lunch*
13:00 - 13:30	*Machine Translation for Highly Low-Resource Language: A Case Study of Ainu, a Critically Endangered Indigenous Language in Northern Japan*
13:30 - 14:00	*Unlocking Transitional Chinese: Word Segmentation in Modern Historical Texts*
14:00 - 14:30	*Combating Hallucination and Misinformation: Factual Information Generation with Tokenized Generative Transformer*
14:30 - 14:45	*Coffee break*
14:45 - 15:15	*A Quantitative Discourse Analysis of Asian Workers in the US Historical Newspapers*
15:15 - 15:45	*Measuring the distribution of Hume's Scotticisms in the ECCO collection*
15:45 - 16:15	*Effect of data quality on the automated identification of register features in Eighteenth Century Collections Online*
16:15 - 16:45	*Comparing Transformer and Dictionary-based Sentiment Models for Literary Texts: Hemingway as a Case-study*

Emotion-based Morality in Tagalog and English Scenarios (EMoTES-3K): A Parallel Corpus for Explaining (Im)morality of Actions

Jasper Kyle Catapang
De La Salle-College of Saint Benilde
Manila City, Philippines
jasperkyle.catapang@benilde.edu.ph

Moses Visperas
University of the Philippines Diliman
Quezon City, Philippines
moses.visperas@eee.upd.edu.ph

Abstract

Grasping morality is vital in AI systems, particularly as they become more prevalent in human-focused applications. Yet, research is scarce on this topic. This study presents the Emotion-based Morality in Tagalog and English Scenarios (EMoTES-3K), a collection that shows commonsense morality in both Filipino and English. This dataset is instrumental for analyzing moral decisions in various situations and their justifications. Our tests show that EMoTES-3K is effective for moral text categorization, with the fine-tuned RoBERTa model scoring 94.95% accuracy in English and 88.53% in Filipino. The dataset also excels in text generation tasks, as shown by fine-tuning the FLAN-T5 model to produce clear moral explanations. However, the model faces challenges when dealing with actions that have mixed moral implications. This work not only bridges the gap in moral reasoning datasets for languages like Filipino but also sets the stage for future research in commonsense moral reasoning in artificial intelligence.

1 Introduction

Moral reasoning, a cornerstone of human cognition, allows individuals to discern right from wrong and make judgments grounded in ethical considerations. As the integration of artificial intelligence (AI) systems into our daily lives deepens, the imperative for these systems to comprehend and reason about moral dilemmas becomes increasingly pronounced. The challenge lies not just in teaching machines to mimic human moral judgments but in ensuring that these judgments are grounded in a robust understanding of ethical principles. This paper aims to bridge a significant gap in the field: the absence of moral reasoning datasets for low-resource languages such as Filipino. Specifically, we introduce a parallel corpus for commonsense morality—determined heavily by emotions—available in both Filipino and English and analyze its validity by using it in downstream tasks, namely text classification and text generation. We call this corpus the Emotion-based Morality in Tagalog and English Scenarios corpus or EMoTES-3K.

The following are the contributions of the researchers:

1. Introduce a commonsense morality dataset in Filipino and English.
2. Demonstrate the dataset's utility in moral text classification and text generation.
3. Demonstrate the (in)ability of large language models to generalize to tricky scenarios in explaining commonsense morality.

2 Background

2.1 Moral Reasoning Frameworks

Jiang et al. (2021) introduced Delphi, an AI system for commonsense moral reasoning. At its core is the Commonsense Norm Bank with 1.7M crowd-sourced ethical judgments. A notable subset is the ETHICS dataset (Hendrycks et al., 2021), covering diverse moral concepts. Building on this, Pyatkin et al. (2023) presented CLARIFYDELPHI, which emphasizes the importance of context in moral reasoning. In parallel, Zhou et al. (2023) suggested rethinking machine ethics with a top-down approach rooted in established moral theories, aiming for greater transparency in AI decision-making.

2.2 Challenges in Moral Enhancement of AI

Understanding the complexities involved in AI's moral reasoning leads us to explore the associated challenges. Serafimova (2020) discussed the differences between moral agency in humans and AI, highlighting the issue of replicating human moral autonomy in machines. The study also underlined the risks of biases in algorithm design, which can lead to computational and moral errors, affecting AI's moral outcomes and raising significant socio-political concerns.

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 1–6
December 1-3, 2023 ©2023 Association for Computational Linguistics

2.3 Human Values and AI Alignment

The relationship between human values and AI decision-making is further elucidated by Sorensen et al. (2023). They introduced VALUE PRISM, a dataset capturing human values in authored situations, and KALEIDO, a model adept at generating and assessing the relevance of these values. Additionally, Yao et al. (2023) focused on aligning Large Language Models (LLMs) with human values, highlighting the evolution of LLMs from basic capabilities to a deep value orientation. Complementing this, Schramowski et al. (2022) showed that LLMs can represent moral norms geometrically and learn moral biases, suggesting their potential in answering moral questions.

2.4 Emphasis on Moral Judgment in Tagalog and Taglish

Transitioning to the linguistic aspect of moral reasoning, our research uniquely focuses on moral judgments in Tagalog and Taglish. These languages play a significant role in the global linguistic landscape, with Tagalog offering a distinctive cultural and moral framework and Taglish providing a blend of local and global moral perspectives. This dual focus allows us to analyze moral reasoning in two linguistically and culturally intertwined environments, highlighting the dynamic interplay of language and culture in moral reasoning.

2.5 Overview of Filipino Morality

To further understand the cultural context of our research, we delve into Filipino morality. Rooted in its cultural, historical, and sociological fabric, Filipino morality integrates indigenous values, colonial influences, and modern global perspectives (Jocano, 1997; Mercado, 1974). The concept of 'kapwa' or shared identity is central to this ethos, emphasizing community and empathy (Enriquez, 2013). The influence of Catholicism and the Philippines' colonial history have shaped a resilient and adaptable moral framework (Constantino, 2022; Doeppers, 2016), which is crucial for developing culturally sensitive AI systems.

2.6 Morphological Challenge in Natural Language Understanding of Filipino

Finally, we address a specific linguistic challenge in the Filipino language. The complex morphology of Filipino verbs (De Guzman, 1978) poses a unique challenge for AI-based moral judgment. Unlike English, with its simpler verb structure, Filipino verbs undergo significant morphological changes that affect moral implications. For example, the verb 'gawa' (to do) in its root form is neutral, but when transformed into 'magagawa' (can do), it implies capability or potential, introducing moral considerations like responsibility and choice. Conversely, 'nagawa' (did) indicates completed action, shifting the focus to accountability for actions taken. Beyond verb forms, Filipino's reliance on context sensitivity, non-verbal cues, and indirect communication style further complicates AI's interpretation of moral nuances. The prevalent use of Taglish, blending Tagalog and English, adds another layer of complexity, reflecting cultural intermingling but posing challenges for AI models trained on monolingual datasets. Specialized algorithms are required to navigate these nuances, understanding the subtle shifts in meaning and cultural implications inherent in Filipino verb forms and communication styles.

3 Experimental Setup

3.1 Dataset Creation

In developing the EMoTES dataset, we meticulously followed an annotation process inspired by the methodology used in Hendrycks et al. (2021). Our team of annotators consisted of bilingual Filipino college graduates with specialized backgrounds in philosophy, psychology, and linguistics. This diverse academic expertise was crucial in ensuring a deep and accurate interpretation of the dataset.

Adopting a quality control approach parallel to that of Hendrycks et al. (2021), each entry in our dataset was subjected to multiple reviews. This rigorous process aimed to ensure consistency in annotations and minimize ambiguities. However, it is essential to acknowledge a potential limitation: all our annotators were from Metro Manila. This geographical concentration may introduce a cultural bias, potentially limiting the representation of diverse provincial values and norms in the Philippines. Therefore, we advise caution when applying our findings to broader, more culturally varied contexts.

Building upon the foundational work of Hendrycks et al. (2021), our research included the creation of approximately 2,400 original scenarios where commonsense moral judgments determine morality. Additionally, from the work of Hendrycks et al. (2021), we adapted and translated about 500 examples—providing not only classifications but also explanations and inferred personality traits from each scenario. In total, the EMoTES-3K dataset comprises 1,712 moral scenarios and 1,193 immoral scenarios.

To illustrate, consider this example from the dataset:

> `Filipino:` "Si Sofia ay nagbebenta ng pekeng COVID-19 test results upang makalusot sa mga travel restrictions."
> `English:` "Sofia is selling fake COVID-19 test results to bypass travel restrictions."
> `Annotation:` Immoral
> `Reason:` "Sofia's action of selling fake COVID-19 test results to evade travel restrictions is highly immoral as it compromises public safety and deceives authorities."
> `Personality Traits:` "deceptive"

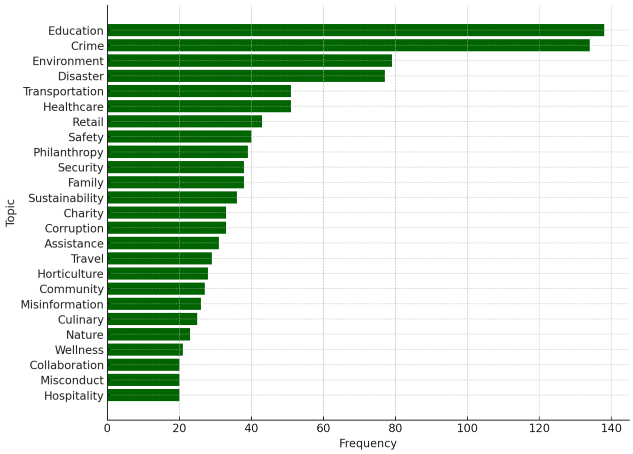


Figure 1: Top 25 topics in EMoTES-3K

The EMoTES-3K dataset explores a diverse array of topics, reflecting a wide spectrum of emotional and thematic elements in textual data. As shown in Figure 1, 'Education' and 'Crime' are the most prominent topics, each with over 130 instances, signifying a strong focus on these societal aspects. Other significant topics include 'Environment', 'Disaster', and 'Transportation', collectively addressing global challenges and daily life concerns. The inclusion of topics such as 'Healthcare', 'Retail', and 'Safety' highlights the dataset's relevance to public welfare and economic activities. Furthermore, the presence of subjects like 'Horticulture', 'Misinformation', and 'Hospitality' provides insights into specialized areas. This assortment of topics in the EMoTES-3K dataset ensures comprehensive coverage of societal themes and underscores its utility in analyzing the complex interplay between topic content and emotional expression. This variety is vital for the development of robust natural language processing tools capable of discerning the nuanced relationships between topics and emotional expressions.

Each scenario within the dataset has an average word count of 12.65 words for English and 15.41 words for Filipino. The distribution of the most common words for each language scenario is illustrated in Figure 2.

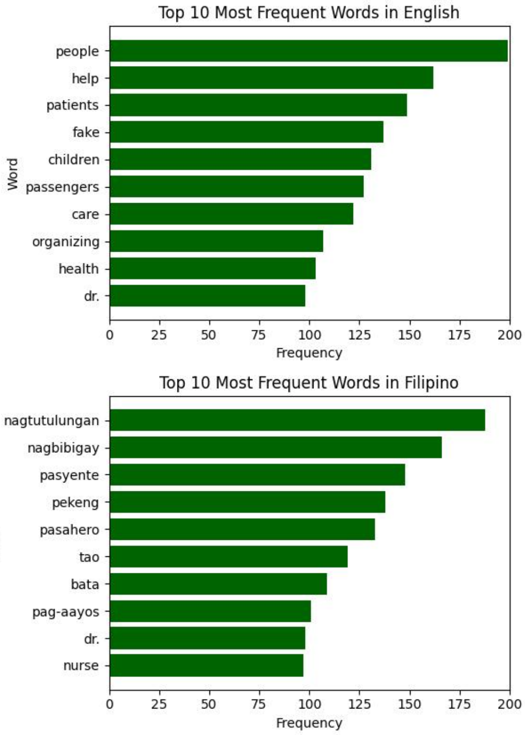


Figure 2: Common words in EMoTES-3K per language

To explain the structure clearly, the data fields of EMoTES-3K are described in Table 1.

Field	Description
entry_id	unique identifier
Filipino	scenario in Filipino
English	scenario in English
Annotation	Moral/Immoral
Explanation	why action is moral or immoral
Personality Traits	inferred traits from action

Table 1: Description of the proposed dataset fields

3.2 Moral Text Classification

One possible application of the EMoTES-3K dataset is the classification of a text's commonsense morality. To show how the corpus can be used for such a task, we use the RoBERTa architecture (Liu et al., 2019). Assuming Schramowski et al. (2022)'s findings hold, language models, like RoBERTa, should yield favorable results. For both language subsets of EMoTES-3K, we use Google Colab's free-tier GPU T4 runtime to fine-tune RoBERTa. No text preprocessing is made.

In the training phase for the RoBERTa model on the English dataset, several hyperparameters are meticulously chosen to ensure optimal performance. The batch size for both training and evaluation is set to 64. The model is trained for a total of 30 epochs. A learning rate of 1×10^{-6} is employed, accompanied by a weight decay of 0.005 to prevent overfitting. The evaluation strategy is configured to evaluate at regular step intervals, specifically every 100 steps, which is also the frequency at which the model's performance metrics were logged. The metrics in consideration are training loss, evaluation loss, accuracy, and F1 score.

3.3 Text Generation

Using EMoTES-3K, we fine-tune the FLAN-T5 large model (Chung et al., 2022) for text generation in both English and Filipino subsets. The prefix "*Explain the morality of this scenario*" was appended to each scenario in the dataset, following FLAN-T5's instruction style training. The model is trained over 30 epochs with a learning rate of 1×10^{-4}, using a linear learning rate scheduler. We set batch sizes at 2 for both training and evaluation and utilize a 32GB V100 GPU from DOST-ASTI's COARE for training[1]. This fine-tuning adapts FLAN-T5 to the EMoTES-3K dataset's specifics, optimizing its performance for subsequent tasks.

The evaluation metrics for this task are ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum, and METEOR. While ROUGE (Lin, 2004) and METEOR(Banerjee and Lavie, 2005) are primarily designed for assessing the quality of machine-generated text in the context of summarization and machine translation, we employ them as a proxy metric because our primary aim is to ensure that our model's outputs align with the desired standard of coherence and relevance to our gold explanations.

4 Results and Discussion

4.1 RoBERTa Sequence Classification

code	train loss	val loss	acc	f1
en	0.076	0.154	0.950	0.959
tl	0.240	0.251	0.885	0.907

Table 2: Training and validation results of RoBERTa on the English and Filipino scenarios at different steps.

RoBERTa model's training and validation results for both English and Filipino datasets are detailed in Table 2, respectively. Two separate models were fine-tuned [2]. For the English dataset, the model's peak performance was at step 800 with an accuracy of 94.95% and an F1 score of 0.9586. In contrast, the Filipino dataset saw its best results at step 900, achieving an accuracy of 88.53% and an F1 score of 0.9070.

The model's stronger performance in English can be attributed to its inherent design for the English language. Additionally, the Filipino language's complexity, marked by its diverse verb inflections, poses challenges in discerning moral implications, making it a more intricate task than in English.

4.2 FLAN-T5 Text Generation

4.2.1 Inherently (Im)moral Scenarios

The training and validation loss of two fine-tuned FLAN-T5 large models are shown in Table 3. After 30 epochs, the English model achieved a training loss of 0.0001 and a validation loss of 0.0546, while the Filipino model reached a training loss of 0.0001 and a validation loss of 0.0711.

Metric	English	Filipino
Training Loss	0.0001	0.0546
Validation Loss	0.0001	0.0711
ROUGE-1	0.7601	0.7263
ROUGE-2	0.7084	0.6562
ROUGE-L	0.7513	0.7146
ROUGE-Lsum	0.7515	0.7142
METEOR	0.8677	0.8249

Table 3: Performance metrics of fine-tuning FLAN-T5 on the text generation task via EMoTES-3K dataset.

Table 3 displays the ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum, and METEOR scores for both the English and Filipino datasets. The

[1] DOST-ASTI COARE's website: https://asti.dost.gov.ph/projects/coare/

[2] Data and fine-tuned models can be found here

ROUGE scores represent the count of matching n-grams, including unigrams, bigrams, and the longest common subsequence, between output and its corresponding reference. Meanwhile, the METEOR metric is based on the precision and recall of matching unigrams. It is evident that our fine-tuned models exhibit exceptional text generation capabilities, as most of our ROUGE metrics achieve scores of 0.70 or higher. The only exception is ROUGE-2 for Filipino. The METEOR scores were also higher for English than Filipino with a score of 0.8677 compared to 0.8249. In general, when we compare text generation capabilities between English and Filipino, we observe that Filipino falls short, likely due to the linguistic intricacies associated with this language, as mentioned earlier.

4.2.2 Scenarios with Mixed Moral Implications

To demonstrate how much the fine-tuned FLAN-T5 understands commonsense morality, we qualitatively analyze two scenarios wherein a moral action is done with ill intent (MB) and two other scenarios where an immoral action is done with good intent (IG). The examples are as follows:

1. MB: “Sarah helps her younger brother with his homework because she wants to show her parents that she’s the better sibling.”
2. MB: “Si John ay nagbigay ng malaking pera sa isang charity para matabunan ang masasamang balita sa kanya.”
   (TRANSLATED): “John gave a huge sum of money to a charity to get rid of the bad news about him.”
3. IG: “Mike doesn’t share his dad’s terminal condition with his mom because she will have a heart attack when she receives the news.”
4. IG: “Ninakaw ni Laura ang susi sa kotse ng kanyang lolang may malabong mga mata upang hindi magamit ng lola niya ang kotse.”
   (TRANSLATED): “Laura stole the car keys of her grandma with poor eyesight so that her grandma can’t use the car.”

The fine-tuned FLAN-T5 models gave the following responses:

1. “This action is considered moral because helping one’s younger sibling with their homework is supportive and educational.”
2. “This action is considered **moral** because donating money to a charity with the intention of embezzling money from it is a highly **immoral** and illegal act.”
3. “Mike’s action of not sharing his dad’s terminal condition to his mom is immoral as it puts her life in immediate danger and puts her health at risk.”
4. “Laura’s action of stealing her grandma’s car **seat** is immoral as it is disrespectful and inconveniences her family members.”

In Example 1, it can be inferred that the model only focused on the action itself and excluded the reason, so the model considered the person’s action as moral. In Example 2, it can be inferred that the model is confused since it was inconsistent in determining whether the action is moral or immoral. In Example 3, the model does not seem to know why the action is immoral, i.e., the judgment is correct but the reasoning is not. Lastly, in Example 4, the model changed the keys to a seat but chose to consider the act of stealing immoral as it is disrespectful.

5 Conclusion

This research introduced the corpus: Emotion-based Morality in Tagalog and English Scenarios or EMoTES-3K, providing a moral reasoning dataset for the Filipino language. The dataset was validated through text classification and text generation tasks. Notably, the RoBERTa model achieved an accuracy of 94.95% for English and 88.53% for Filipino in moral text classification. Furthermore, the fine-tuned FLAN-T5 models showcased impressive text generation capabilities, with most ROUGE metrics surpassing 0.70. However, the models exhibited challenges in discerning complex moral scenarios, especially those with conflicting intents. This study not only fills a gap in moral reasoning datasets for languages like Filipino but also highlights the intricacies and challenges of teaching AI systems to reason about morality.

Acknowledgements

We extend our gratitude to the Advanced Science and Technology Institute (ASTI) and the Computing and Archiving Research Environment (COARE) for providing our team with access to their High-Performance Computing (HPC) resources. We are also thankful to the manuscript reviewers and dataset annotators, whose insights and efforts were essential in elevating the quality of our research.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Renato Constantino. 2022. The philippines: A past revisited.

Videa P De Guzman. 1978. Syntactic derivation of tagalog verbs. *Oceanic linguistics special publications*, (16):i–414.

Daniel F Doeppers. 2016. *Feeding Manila in peace and war, 1850–1945*. University of Wisconsin Pres.

Virgilio Enriquez. 2013. From colonial to liberation psychology: The philippine experience. *Philosophy East and West*, 63(2).

Dan Hendrycks, Andrew Critch, Collin Burns, Jerry Li, Dawn Song, Steven Basart, and Jacob Steinhardt. 2021. Aligning ai with shared human values. *OpenReview*.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, et al. 2021. Can machines learn morality? the delphi experiment. *arXiv preprint*.

F Landa Jocano. 1997. Filipino value system: a cultureal definition. *(No Title)*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Leonardo N Mercado. 1974. Elements of filipino philosophy.

Valentina Pyatkin, Jena D. Hwang, Vivek Srikumar, Ximing Lu, Liwei Jiang, Yejin Choi, and Chandra Bhagavatula. 2023. Clarifydelphi: Reinforced clarification questions with defeasibility rewards for social and moral situations. In *ACL Anthology*.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain humanlike biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268.

Silviya Serafimova. 2020. Whose morality? which rationality? challenging artificial intelligence as a remedy for the lack of moral enhancement. *Humanities and Social Sciences Communications*, 7:119.

Taylor Sorensen, Liwei Jiang, Jena D. Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. 2023. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. *arXiv preprint*.

Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. 2023. From instructions to intrinsic human values — a survey of alignment goals for big models. *arXiv preprint*.

Jingyan Zhou, Minda Hu, Junan Li, Xiaoying Zhang, Xixin Wu, Irwin King, and Helen Meng. 2023. Rethinking machine ethics – can llms perform moral reasoning through the lens of moral theories? *arXiv preprint*.

A Quantitative Discourse Analysis of Asian Workers in the US Historical Newspapers

Jaihyun Park[1], **Ryan Cordell**[1]
[1]School of Information Sciences
[1]University of Illinois at Urbana-Champaign
jaihyun2@illinois.edu, rcordelll@illinois.edu

Abstract

Warning: This paper contains examples of offensive language targetting marginalized population.
The digitization of historical texts invites researchers to explore the large-scale corpus of historical texts with computational methods. In this study, we present computational text analysis on a relatively understudied topic of how Asian workers are represented in historical newspapers in the United States. We found that the word "coolie" was semantically different in some States (e.g., Massachusetts, Rhode Island, Wyoming, Oklahoma, and Arkansas) with the different discourses around coolie. We also found that then-Confederate newspapers and then-Union newspapers formed distinctive discourses by measuring over-represented words. Newspapers from then-Confederate States associated coolie with slavery-related words. In addition, we found Asians were perceived to be inferior to European immigrants and subjected to the target of racism. This study contributes to supplementing the qualitative analysis of racism in the United States with quantitative discourse analysis.

1 Introduction

Digitization of historical texts has opened up new opportunities for researchers to explore a large-scale corpus of historical texts with computational methods. Especially, the ready availability of datasets required for Natural Language Processing (NLP) research has welcomed researchers from other disciplines to NLP research (Park and Jeoung, 2022) and reduced the barriers to entry of NLP research. Taking this advantage, many researchers in diverse field, namely Sociology, History, English, and Information Science have applied NLP techniques to historical texts, such as books (Parulian et al., 2022), newspapers (Smith et al., 2013; Pedrazzini and McGillivray, 2022; Santin et al., 2016), and/or congressional records (Lin and Peng, 2022; Guldi, 2019) by the help of available archival metadata (Dobreski et al., 2019). Creating an interdisciplinary research space called *Digital Humanities*, there have been studies primarily focusing on the race problem in the United States. To introduce a few large-scale computational research, Soni et al. (2021) traced the semantic change of the word, for instance, when and which newspaper started to use the word with a new meaning and when and which newspapers adopted new semantic meaning of the word in abolitionist newspapers from the 19th century. Franzosi et al. (2012) analyzed racial violence, especially lynching performed by the White mob in the 19th century Georgia and presented quantitative narratives of the racial violence. These studies intersect the problem of historical racism and NLP research. However, despite the increasing attention toward the large scale text analysis on historical racism, there has been a gap of understanding the racism posed toward Asians in the United States. It is true that the major racial tension in the United States has been between White and Black. At the center of secession of the South and the creation of Confederacy, there was slavery problem. However, the racial tension between White and Asian has also been a part of the history of the United States. To explore understudied topic of how Asian population was discriminated in the history of the United States, we present computational text analysis on how Asian workers are represented in the U.S. newspapers by searching the derogatory word ("coolie") referencing to Asian workers. We further developed research questions as follows:

- RQ 1. How different are the semantic meaning of "coolie" in each State?
- RQ 2. What are the words over-represented in the newspapers between then-Confederate States and then-Union States?
- RQ 3. What "coolie" stories are reprinted and what are their characteristics?

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 7–15
December 1-3, 2023 ©2023 Association for Computational Linguistics

To support open science and transparent data science, we publish the code used in this study at https://github.com/park-jay/coolie.

2 Background

Throughout 19th and 20th century, Asian immigrant workers were derogatorily called *"coolies."* Britannica entry on *"coolie"* introduces the word as "pejorative European usage" to describe "an unskilled labourer or porter usually in or from the Far East hired for low or subsistence wages." [1] Breman and Daniel (1992) studied the origin of the word "coolie" and claimed the transformation of the word "coolie" from "kuli" (a type of payment for menial work in Tamil) signifies the change of the word from a neutral term to a derogatory term and reflects the person collapse into the payment for labor in English.

In seeking to fill the labor shortage in the United States due to the abolition, Chinese workers were recruited and the migration of Chinese workers arrived in the United States beyond China-neighboring countries like India and Malaysia (Farley, 1968). Even though Asian coolies were perceived to be patient, tractable, obedient, industrious, and frugal compared to African slaves (Jung, 2006), coolies were distrusted, detested, and discriminated (Breman, 2023).

With influx in the number of Chinese workers in the United States, it has come to public's attention that indentured laborers were analogous to modern trafficking and they are no different from slavery (Kempadoo, 2017). Jung (2006) argued that the United States minister to China, William B. Reed viewed coolie trade more than coercion and perceived the coolie problem with the racial and national interest. Reed thought Chinese "would either amalgamate with with the negro race, and thus increase the actual slave population." Rising anti-Chinese sentiment and perceiving as a threat to White workers as their cheap replacement (Rhoads, 2002) and inferior, Chinese Exclusion Act of 1882 remarks the watershed of America's gatekeeping and defining the desirability (and "Whiteness") of immigrant groups (Lee, 2002).

The problem of coolie exemplifies the extension of colonial and capitalist exploitation beyond Africa and sugarcoated the extended system as indentured migrant contract workers (Van Rossum, 2016). Therefore, studying coolie problem in the United States context supplements the historical study of racial tension mostly focused between White and Black and extends the racial conflict to include Asian population.

3 Methodology

3.1 Data collection

The data is collected from Chronicling America [2] API where digitized texts through optical character recognition (OCR) are accessible. When the word "coolie" was queried, API showed 124,511 pages of newspapers containing the word [3]. First, we collected the entire pages of the newspapers containing the word "coolie." Then, we extracted the text from the pages and searched the exact match for the word "coolie." This additional step ensures excluding false positive cases due to mis-recognized words. For instance, the search included the result of "cooli" even though it was not identical keyword that we wanted to query. In *New York Daily Tribune* published on August 5th, 1862 contained the word "cooli" but it accompanied many OCR errors making it doubtful whether the word "cooli" was actually from the word "coolie." [4] In order to reduce this kind of false positive cases, we double-processed the data by finding the exact match for the word "coolie" in the extracted text. Instead of finding a sentence that contains the word "coolie", we extracted upto ten tokens before the word "coolie" appeared and upto ten tokens after the word "coolie" appeared to create a pseudo-sentence. As some digitization of the newspapers were not perfect and thus missed punctuation marks, sentence tokenizer could not identify the sentence boundary correctly. This additional step of creating pseudo-sentence resulted in 125,253 text data (pseudo-sentence) containing the word "coolie" for the analysis. The earliest publication date was June 30th, 1795 and the latest publication date was December 6th, 1963.

In figure 1, we present the count of text data containing the word "coolie" by State. The count of text data is not evenly distributed across the States due to different digitization process of the newspapers. The most text data was from Dis-

[1] https://www.britannica.com/money/topic/coolie-Asian-labourer

[2] https://chroniclingamerica.loc.gov/

[3] The data is collected on September 5th, 2023.

[4] https://chroniclingamerica.loc.gov/lccn/sn83030213/1862-08-05/ed-1/seq-4/ocr/ was searched as a page that contains "coolie" but it was due to the OCR error.

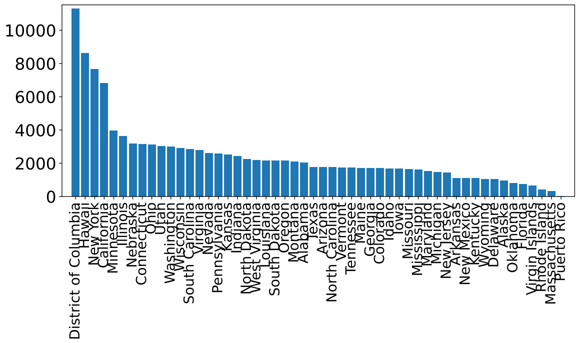


Figure 1: The count of text data containing the word "coolie" by State

trict of Columbia (*n=11,302*) followed by Hawaii (*n=8,613*) and New York (*n=7,671*). Puerto Rico (*n=15*), Massachusetts (*n=305*), Rhode Island (*n=418*), and Virgin Islands (*n=656*) had the least text data.

3.2 Data pre-processing

We pre-processed the data by removing the punctuation, non-alphabet tokens that might have been mis-recognized during the OCR process, and stop words. The list of stop words is from NLTK package in Python and we further converted the words into lemmas using Spacy [5]. Before we feed the data into the word embedding model, we ran the FastText model (Bojanowski et al., 2017) to identify possible OCR errors. FastText embedding takes character n-gram as input and outputs the embedding vector. This model was tested effective that it can generate possible OCR error candidates (Hajiali et al., 2022). By training entire sentence that contains the word "coolie", we identified 200 most similar words to the word "coolie". For instance, "coolieize" (0.8654), "oroolie" (0.8630), and "roolie" (0.8541) ranked high in the list of similar words to "coolie" according to the FastText embedding model. With 200 most similar words, we changed the top 200 words in the text data into "coolie" and trained the Word2vec model in section 3.3.

3.3 RQ1. Word embedding

In order to answer RQ 1, we trained the Word2vec model (Mikolov et al., 2013) to use Continuous Bags of Words (CBOW) approach and the skip-gram approach. Both CBOW and skip-gram approach find the word embedding by predicting the target word from the context words. We trained the Word2vec model with minimum word count of 5 and window size of 5 to generate the word embedding. We then took the average of the word embedding vector of the word "coolie" in each state and calculated the cosine similarity.

3.4 RQ2. Statistically over-represented words

In answering RQ 2, we grouped the newspapers into two groups: the newspapers from the then-Confederate States and then-Union newspapers. For the then-Confederate States, we included the newspapers from Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Texas, and Virginia. For the then-Union States, Maine, New York, New Hampshire, Vermont, Massachusetts, Connecticut, Rhode Island, Pennsylvania, New Jersey, Ohio, Indiana, Illinois, Kansas, Michigan, Minnesota, Wisconsin, Iowa, California, Nevada, Oregon, Delaware, Maryland, and West Virginia were included. We excluded sentences from the newspapers newspapers located in Virgin Islands (*n=656*) and Puerto Rico (*n=15*).

We calculated the log-odds ratio with informative Dirichlet prior (Monroe et al., 2008) by comparing the word frequency in the then-Confederacy newspapers and the newspapers published in the rest of the United States. The detailed metric is provided in equation 1.

$$\begin{aligned}\delta_w^{(i-j)} = \log & \frac{y_w^i + a_w}{n^i + a_0 + y_w^i - a_w} \\ - \log & \frac{y_w^i + a_w}{n^i + a_0 - y_w^i - a_w}\end{aligned} \tag{1}$$

The log-odds ratio with informative Dirichlet of each word w between two corpora i and j (in our study, newspapers from the then-Confederate States and the rest) given the prior frequencies are obtained from the entire corpus a. We selected 15,000 most frequent words from the entire corpus and the Z-score is calculated for each word. When n^i is the total number of words in corpus i, y_w^i is the number of times word w appears in corpus i, a_0 is the size of the corpus a, and a_w is the frequency of word w in corpus a (Kwak et al., 2020). With the log-odds ratio, we can identify the words that are over-represented in the corpora.

3.5 RQ 3. Text reprint detection

Nineteenth-century American newspapers reprinted texts from a wide range of genres:

[5] https://spacy.io/

news reports, recipes, trivia, lists, vignettes, and religious reflections (Cordell and Mullen, 2017). Text reprints could also include boilerplate that appeared across many issues of the same paper, such as advertisements. A business might buy ad space for multiple weeks, months, or even years, and those ads would be left in standing type from issue to issue. In a study such as this one, focused on textual reuse, an ad that includes a keyword of interest but which appears day after day can disproportionately influence the statistical relationship between words in the corpus, leading our model to overestimate the importance of words within the ad relative to the words in texts that changed each day. In other words, if one particular phrase repeatedly appears, then the embedding model will overfit the phrase because of the distorted distribution of the text. However, it is hard to detect reprints based on keyword searches because of OCR errors. Here we adopt the Viral Texts project's text-reuse detection methods, as described in Smith et al. (2014) , which use n-gram document representations to detect text reprints within errorful OCR-derived text. We processed our corpus with a 5-gram chunking using NLTK whitespace tokenizer and further made a judgment that the text has been reprinted when there were more than three matches of 5-grams across the snippets. With this method, the negative impact of the OCR errors can be reduced. For instance, this pair: ["demolish", "part", "build", "injure", "two", "coolie", "police", "investigation", "latter", "case", "lead"] and ["demolish", "part", "build", "jure", "two", "latter" "case", "lead"] are not identical because of inconsistent OCR like "injure" and "jure". However, the 5-gram matching examination substantiated that this pair denotes a reprinted text. Due to the effectiveness of the method by Smith et al. (2014), we used the method to answer RQ 3.

4 Results

4.1 RQ1. Comparing the meaning of "coolie"

In figure 2, we present the heatmap of cosine similarity comparison across the average embedding vector of the word "coolie" in each state. We are visually informed that the cosine similarity in most of the States is high except for a few States that created a semantically different meaning of the word "coolie". For instance, Massachusetts and Rhode Island showed average cosine similarity of 0.08 and 0.12, respectively when average cosine similarity across the entire States was 0.65. The most semantically dissimilar State to Massachusetts was when compared to Oklahoma (-0.10) and the most similar State to Massachusetts was North Dakota (0.23). In the meantime, Rhode Island showed the most dissimilar meaning of the word "coolie" when compared to Delaware (-0.03) and the most similar State to Rhode Island was Mississippi (0.23). Some then-Confederate States showed lower cosine similarity than the average cosine similarity across the entire States. For instance, Arkansas (0.43), Florida (0.48), and Tennessee (0.63) showed lower cosine similarity on average. The most dissimilar State to Arkansas was Massachusetts (-0.06) and the most similar State to Arkansas was Colorado (0.54). For Florida, the most dissimilar State was Rhode Island (0.06) and the most similar State to Florida was Nevada and Utah (0.61). For Tennessee, the most dissimilar State was Massachusetts (-0.03) and the most similar State to Tennessee was Utah (0.80).

The highest five States				
Illinois	California	Wisconsin	Virginia	Nevada
labor (0.9998)	country (0.9995)	chinese (0.9998)	chinese (0.9998)	chinese (0.9997)
chinese (0.9998)	bill (0.9995)	labor (0.9998)	man (0.9997)	club (0.9997)
wage (0.9997)	upon (0.9995)	two (0.9998)	trade (0.9997)	labor (0.9997)
two (0.9997)	well (0.9995)	one (0.9998)	work (0.9997)	make (0.9997)
one (0.9997)	go (0.9995)	time (0.9998)	one (0.9997)	use (0.9996)
day (0.9997)	stop (0.9995)	carry (0.9997)	three (0.9997)	say (0.9996)
china (0.9997)	many (0.9995)	take (0.9997)	number (0.9997)	importation (0.9996)
man (0.9997)	american (0.9995)	make (0.9997)	make (0.9997)	man (0.9996)
pay (0.9997)	con (0.9995)	japanese (0.9997)	importation (0.9997)	trade (0.9996)
say (0.9997)	would (0.9995)	would (0.9997)	two (0.9997)	day (0.9996)

Table 1: Top 10 most similar words to the word "coolie" in the top 5 States that showed the most similar meaning of the word "coolie"

To delve into the semantic difference of the word "coolie" in each State, we present the top 10 most similar words to the word "coolie" in the top 5 States that showed the most similar meaning of the word "coolie" and the top 5 States that showed the most dissimilar meaning of the word "coolie" in table 1 and table 2.

In table 1, we can observe that the word "coolie" was used in the context of labor, China, and wage.

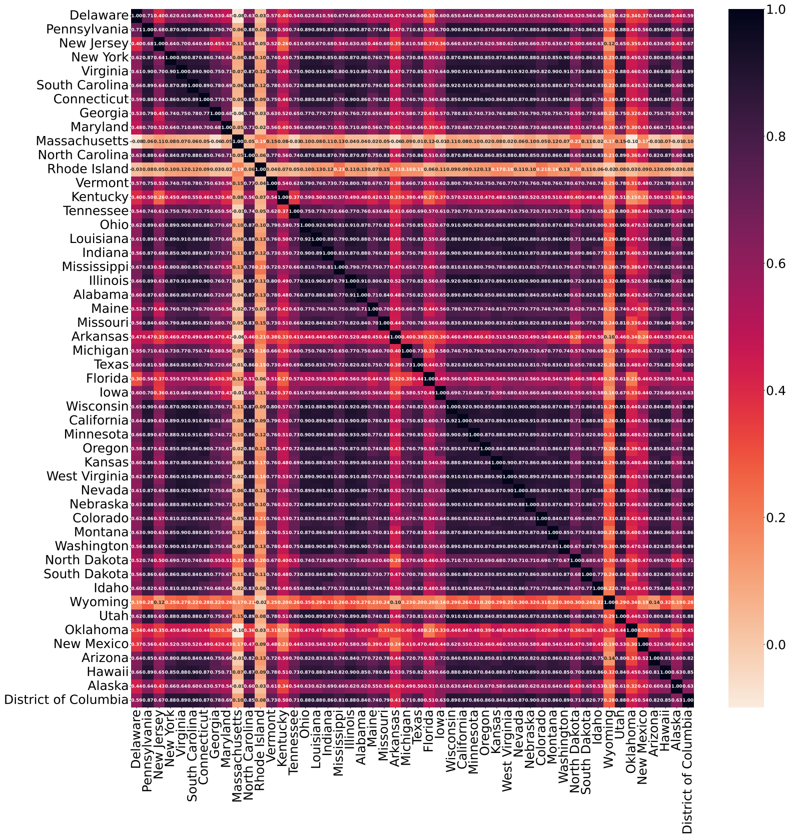


Figure 2: The heatmap of cosine similarity comparison across the average embedding vector of the word "coolie" in each state

The word related to labor ("labor" and "work") appeared in the semantically close words in Illinois, Wisconsin, Virginia, and Nevada. "chinese" was the most similar word of identification of coolie's ethnicity (Illinois, Wisconsin, Virginia, and Nevada) while "japanese" appeared in Wisconsin as well. The word related to wage ("wage" and "pay") appeared in the high closest word to "coolie."

Among the highest five States, there are many common words across the States that might have created an embedding for the word "coolie" not so much different from other States. Words like "make", "say", "man", and/or numbers like "one" and "two" are common words across the highest five States.

The lowest five States				
Massachusetts	Rhode Island	Wyoming	Oklahoma	Arkansas
among (0.3579)	order (0.4116)	chinese (0.9969)	chinese (0.9821)	labor (0.9985)
call (0.3201)	india (0.3972)	labor (0.9969)	shoulder (0.9815)	chinese (0.9984)
know (0.2832)	woman (0.3922)	would (0.9955)	japanese (0.9776)	mongolian (0.9978)
prohibit (0.2295)	great (0.3890)	japanese (0.9952)	pay (0.9748)	one (0.9978)
time (0.2232)	take (0.3761)	six (0.9950)	also (0.9725)	thousand (0.9977)
report (0.2221)	ship (0.3745)	one (0.9949)	labor (0.9706)	japanese (0.9975)
get (0.2209)	law (0.3432)	bring (0.9948)	carry (0.9632)	tolerate (0.9975)
arrive (0.2176)	united (0.3260)	say (0.9946)	home (0.9632)	revivial (0.9974)
come (0.2131)	carry (0.3234)	work (0.9944)	get (0.9630)	may (0.9974)
two (0.2104)	many (0.3122)	two (0.9941)	work (0.9612)	carry (0.9974)

Table 2: Top 10 most similar words to the word "coolie" in the top 5 States that showed the most dissimilar meaning of the word "coolie"

However, the most dissimilar States, Massachusetts and Rhode Island, showed the different context of the word "coolie" as presented in table 2. Unlike the highest five States, where the word "coolie" was used in the context of labor, China, and wage, Massachusetts and Rhode Island created a unique discourse of the word "coolie." For instance, the word "labor" and "work" does not appear in the top 10 most similar words to the word "coolie" in Massachusetts and Rhode Island. Similarly, we cannot find "chinese" or "japanese" that refer to the ethnicity of coolie workers in the top 10 most

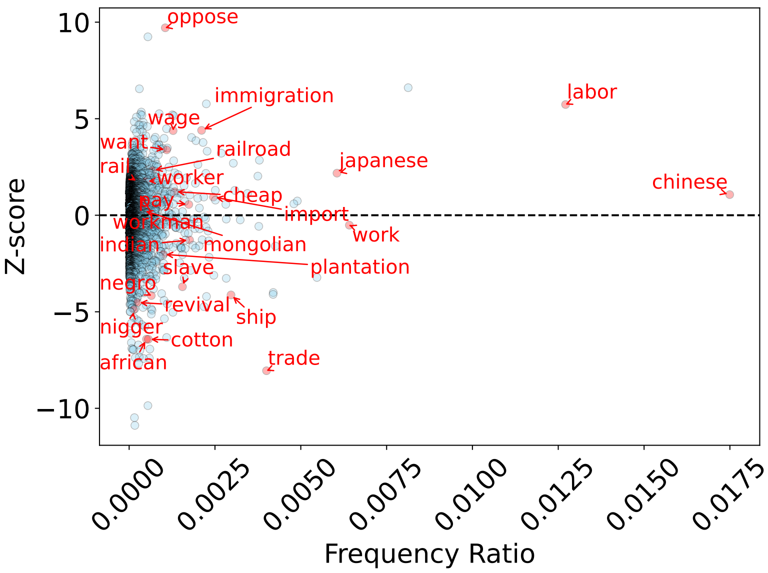


Figure 3: The Z-score of words in then-Confederate and then-Union newspapers

similar words as well as "wage" or "pay." Instead, Massachusetts has "among", "call", "know", "prohibit", "report", "get", "arrive", and "come" as semantically close words to "coolie." However, these words are not found in top 10 closest words in the highest five States in table 1. A unique embedding for the word "coolie" in Rhode Island could also be attributed to the unique discourse created by uncommon words such as "order", "india", "woman", "great", "ship", "law", "united", and "many." However, Wyoming, Oklahoma, and Arkansas share common words with the States in table 1. The words like "chinese", "labor", "would", "japanese", "one", "say", "work" in Wyoming are among top 10 similar words in the highest five States. Despite existence of common words, "shoulder", and "home" in Oklahoma and "mongolian", "tolerate", and "revival" in Arkansas are unique words that are not found in the top 10 similar words in the highest five States.

4.2 RQ 2. Statistically over-represented words in the South and the rest of the United States

We present the result of the log-odds ratio with informative Dirichlet prior in figure 3. The words that are over-represented in the then-Confederate States are located in the area of below 0 while the words that are over-represented in the then-Union States are located in the area of above 0. X-axis represents the frequency ratio of the word in both then-Confederate and then-Union States. Y-axis represents the Z-score of the word. The most frequent word "chinese" have 1.0681 Z-score meaning that the word "chinese" is used relatively more frequently in the then-Union States newspapers than the then-Confederate States newspapers. For "labor", it was over-represented in the then-Union States newspapers with 5.7346 Z-score while semantically similar word "work" was a little skewed to the then-Union States newspapers with -0.5145 Z-score. On the contrary, "worker" and "workman" are over-represented in the then-Union States newspapers with 1.7586 and 1.1122 Z-score, respectively. The word concerning compensation for indentured labor, such as the word "wage" (Z-score=4.3827), "cheap" (Z-score=1.2223), and "pay" (Z-score=0.5588) are over-represented in the then-Union States newspapers. The word about coolie trade ("trade" and "ship") are over-represented in the then-Confederate States newspapers with Z-score of -8.0541 and -4.1275 respectively. The word related to the race whose labor was exploited under the slavery institution, such as "slave" (Z-score=-3.704), "negro" (Z-score=-4.1616), "nigger" (Z-score=-4.8824), and "african" (Z-score=-6.4188) are largely over-represented in the then-Confederate States newspapers. In addition, the location where the labor force was most wanted from coolies were well-captured by log-odds ratio metric. For instance, "plantation" (Z-score=-2.0234) and "cotton" (Z-score=-6.4234) are over-represented in the then-Confederate newspapers while "rail" (Z-score=1.7115) and "railroad"

(Z-score=2.3166) are over-represented in the then-Union newspapers. The finding that the discourse around coolie is associated with slavery-related words in the then-Confederate newspapers supplements historical claim that the indentured Asian laborers were introduced to fill the absence of labor force in the South after abolition of free labor of African Americans (Van Rossum, 2016).

4.3 RQ 3. Reprint network of coolie stories

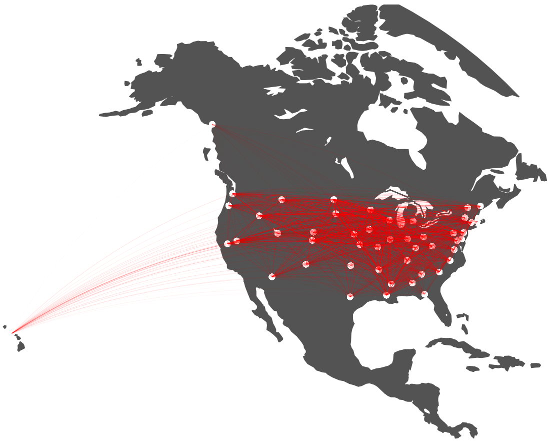
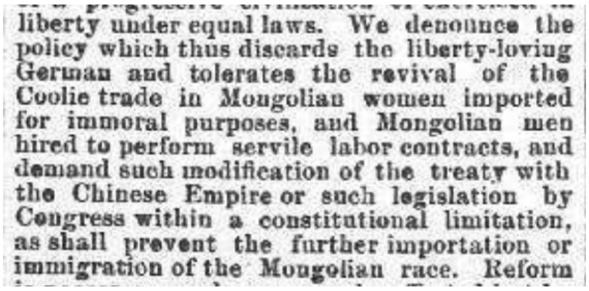


Figure 4: The reprint network of "coolie" stories in the newspapers

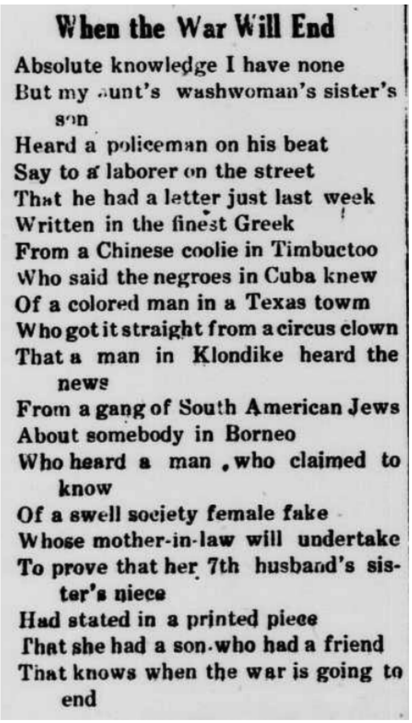
In this section, we present the network of coolie story reprints in figure 4. We represented the Capital city of each State instead of connecting the cities where the newspaper company was located for the sake of brevity of visualization. The network of text reprint about coolie stories shows high average clustering coefficient (0.9905) (Saramäki et al., 2007) due to the presence of reprints spread to multiple States. We found the most reprinted text is from the Democratic party at National Convention. The political message contained the word "coolie." We found 97 reprints of the text from declaration speech from *The Opelousas courier* published in July 8th, 1876.

The excerpt of the message from the Democratic party is presented in figure 5. This statement also implies that Asians (represented as Mongolian with derogatory term) are inferior to "liberty-loving" Germans who could have spread the idea of freedom and solidified the spirit of liberty in the United States that coolies could not bring. Another comparison can be made with treating Asian workers as commodities instead of human beings by calling "coolie trade" and "importation." Indeed, the banning of coolie transportation is based on viewing it as illegal, immoral, and inhuman atrocities resembling African slave trade (Jung, 2006). However, coolies are unwelcoming race compared to European immigrants. Supplementing Lee (2002)'s argument that the desirable quality of immigrants entering the United States was racially White, we observe that there was a prevailing discourse of discriminating Asians against so-called "liberty-loving" Germans.



liberty under equal laws. We denounce the policy which thus discards the liberty-loving German and tolerates the revival of the Coolie trade in Mongolian women imported for immoral purposes, and Mongolian men hired to perform servile labor contracts, and demand such modification of the treaty with the Chinese Empire or such legislation by Congress within a constitutional limitation, as shall prevent the further importation or immigration of the Mongolian race. Reform

Figure 5: The text containing "coolie" in *The Opelousas courier* published in July 8th, 1876



When the War Will End

Absolute knowledge I have none
But my aunt's washwoman's sister's son
Heard a policeman on his beat
Say to a laborer on the street
That he had a letter just last week
Written in the finest Greek
From a Chinese coolie in Timbuctoo
Who said the negroes in Cuba knew
Of a colored man in a Texas towm
Who got it straight from a circus clown
That a man in Klondike heard the news
From a gang of South American Jews
About somebody in Borneo
Who heard a man, who claimed to know
Of a swell society female fake
Whose mother-in-law will undertake
To prove that her 7th husband's sister's niece
Had stated in a printed piece
That she had a son who had a friend
That knows when the war is going to end

Figure 6: The text containing "coolie" in *Middletown transcript* published in April 13th, 1918

The next common reprint in the dataset was from the poem that was circulated through 78 reprints. The poem is presented in figure 6 and it was appeared in *Middletown transcript* published in April

13th, 1918. The overall sentiment in this poem is light and jokey, however, this poem exemplifies the history when racism was naturalized in cultural discourse. This poem borrows marginalized population such as "coolie," "negroes," "colored man," and "jews" which are not necessarily critical components that make rhymes work. Although this poem reflects orientalist view (Said, 1978) by bringing "Timbuctoo," "Greek," and "Klondike " and putting emphasis on exotic sense of information about the war, it can be offensive to people who are settling in the United States from the places mentioned in the poem. The poem emphasizes the distance and exoticism of the places and people who are not White and thus it can be considered as micro aggression toward marginally represented population in the United States.

5 Conclusion

In this study, we present a quantitative discourse analysis on "coolie," a derogatory term referencing to Asian workers which has been understudied in digital humanities field. We used word embedding to compare the meaning of "coolie" in each State and found that Massachusetts, Rhode Island, Wyoming, Oklahoma, and Arkansas showed the most dissimilar meaning of the word "coolie" while Illinois, California, Wisconsin, Virginia, and Nevada showed the most similar meaning of the word "coolie" compared to the rest of the States. We found the reason for the dissimilar meaning of the word "coolie" in Massachusetts and Rhode Island could be in part due to the unique discourse created by uncommon words. With log-odds ratio calculation, we found that the discourse of coolie in the then-Confederate newspapers was accompanied by the words related to African American slavery as well as where the work force is the most needed (e.g., "cotton" and "plantation"). This finding supplements historical argument that Asian workers were introduced to the United States to replace African American labor. With text reuse detection, we found discriminating expression toward Asian workers in political statement and poems that could show stereotype of Asian workers in the United States history.

6 Limitations and Future Work

In addition to data-inherited limitation, OCR errors, digital archives are far from objective and neutral reflection of the past. Digitization of cultural and historical materials is influenced by power politics and it requires a caution in interpreting the results (Zaagsma, 2023) as the result might be grounded in skewed number of available data. For instance, Chronicling America has been criticized for imbalanced number of available data, presenting dominant viewpoints of White compared to small number of digitized Black press (Fagan, 2016). Massachusetts and Rhode Island, the States that showed the most dissimilar meaning of the word "coolie" in our study, are the States that have relatively few number of newspapers containing the word "coolie." As number of available digitized data does not reflect the actual number of newspapers published in the past, we need to be cautious in interpreting the results. In addition, the boundary of the State is not fixed and it is not clear whether the State boundary in the past is the same as the State boundary in the present. With gold rush and railroad construction, the West was rapidly developed and the State boundary was changed. Therefore, after the Civil War, grouping the States into the then-Confederate and then-Union States might not be the best way to group the States as it does not reflect the Westward expansion of the United States. In future work, we will explore more subtle unit of analysis such as geographical location after the civil war to include the rise of the West and how introduction of Chinese Exclusion Act of 1882 changed the meaning of coolie.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146. Publisher: MIT Press.

Jan Breman. 2023. Coolie labour and colonial capitalism in asia. *Journal of Agrarian Change*, 23(2):233–246.

Jan Breman and E Valentine Daniel. 1992. Conclusion: The making of a coolie. *The Journal of Peasant Studies*, 19(3-4):268–295.

Ryan Cordell and Abby Mullen. 2017. " fugitive verses": The circulation of poems in nineteenth-century american newspapers. *American Periodicals*, pages 29–52.

Brian Dobreski, Jaihyun Park, Alicia Leathers, and Jian Qin. 2019. Remodeling archival metadata descriptions for linked archives. In *International Conference on Dublin Core and Metadata Applications*, pages 1–11.

Benjamin Fagan. 2016. Chronicling white america. *American Periodicals: A Journal of History & Criticism*, 26(1):10–13.

M Foster Farley. 1968. The chinese coolie trade 1845-1875. *Journal of Asian and African Studies*, 3(3):257–270.

Roberto Franzosi, Gianluca De Fazio, and Stefania Vicari. 2012. Ways of measuring agency: an application of quantitative narrative analysis to lynchings in georgia (1875–1930). *Sociological Methodology*, 42(1):1–42.

Jo Guldi. 2019. Parliament's debates about infrastructure: an exercise in using dynamic topic models to synthesize historical change. *Technology and Culture*, 60(1):1–33.

Mahdi Hajiali, Jorge Ramón Fonseca Cacho, and Kazem Taghva. 2022. Generating Correction Candidates for OCR Errors using BERT Language Model and FastText SubWord Embeddings. In *Intelligent Computing: Proceedings of the 2021 Computing Conference, Volume 1*, pages 1045–1053. Springer.

Moon-Ho Jung. 2006. *Coolies and cane: Race, labor, and sugar in the age of emancipation*. JHU Press.

Kamala Kempadoo. 2017. 'bound coolies' and other indentured workers in the caribbean: Implications for debates about human trafficking and modern slavery. *Anti-Trafficking Review*, (9):48–63.

Haewoon Kwak, Jisun An, and Yong-Yeol Ahn. 2020. A systematic media frame analysis of 1.5 million new york times articles from 2000 to 2017. In *Proceedings of the 12th ACM Conference on Web Science*, pages 305–314.

Erika Lee. 2002. The chinese exclusion example: Race, immigration, and american gatekeeping, 1882-1924. *Journal of American Ethnic History*, 21(3):36–62.

King Ip Lin and Sabrina Peng. 2022. Enhancing digital history–event discovery via topic modeling and change detection. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 69–78.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26:1–9.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Jaihyun Park and Sullam Jeoung. 2022. Raison d'être of the benchmark dataset: A survey of current practices of benchmark dataset sharing platforms. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 1–10.

Nikolaus Nova Parulian, Ryan Dubnicek, Glen Worthey, Daniel J Evans, John A Walsh, and J Stephen Downie. 2022. Uncovering black fantastic: Piloting a word feature analysis and machine learning approach for genre classification. *Proceedings of the Association for Information Science and Technology*, 59(1):242–250.

Nilo Pedrazzini and Barbara McGillivray. 2022. Machines in the media: semantic change in the lexicon of mechanization in 19th-century british newspapers. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 85–95.

Edward JM Rhoads. 2002. " white labor" vs." coolie labor": The" chinese question" in pennsylvania in the 1870s. *Journal of American Ethnic History*, 21(2):3–32.

Edward W. Said. 1978. *Orientalism*. Pantheon Books.

Bryan Santin, Daniel Murphy, and Matthew Wilkens. 2016. Is or are: The" united states" in nineteenth-century print culture. *American Quarterly*, 68(1):101–124.

Jari Saramäki, Mikko Kivelä, Jukka-Pekka Onnela, Kimmo Kaski, and Janos Kertesz. 2007. Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E*, 75(2):027105.

David A Smith, Ryan Cordel, Elizabeth Maddock Dillon, Nick Stramp, and John Wilkerson. 2014. Detecting and modeling local text reuse. In *IEEE/ACM Joint Conference on Digital Libraries*, pages 183–192. IEEE.

David A Smith, Ryan Cordell, and Elizabeth Maddock Dillon. 2013. Infectious texts: Modeling text reuse in nineteenth-century newspapers. In *2013 IEEE International Conference on Big Data*, pages 86–94. IEEE.

Sandeep Soni, Lauren F Klein, and Jacob Eisenstein. 2021. Abolitionist networks: Modeling language change in nineteenth-century activist newspapers. *Journal of Cultural Analytics*, 6(1).

Matthias Van Rossum. 2016. Coolie transformations–uncovering the changing meaning and labour relations of coolie labour in the dutch empire (18th and 19th century). *Oliver Tappe, Michael Zeuske (eds.) Bonded Labour*, pages 83–103.

Gerben Zaagsma. 2023. Digital history and the politics of digitization. *Digital Scholarship in the Humanities*, 38(2):830–851.

Revisiting Authorship Attribution of *Tirant lo Blanc* Using Parts of Speech *n*-grams

Yoshifumi Kawasaki
University of Tokyo
ykawasaki@g.ecc.u-tokyo.ac.jp

Abstract

Tirant lo Blanc (TLB) is a masterpiece of medieval Catalan chivalric romance. Regarding its authorship, two hypotheses exist: the single-authorship hypothesis claims in agreement with the dedication that Joanot Martorell is the sole author, whereas the dual-authorship hypothesis alleges in line with the colophon that Martorell wrote the first three parts and Martí Joan de Galba added the fourth part. In this study, we revisit the unsettled authorship attribution of TLB with stylometric techniques; specifically, we exploit parts-of-speech (POS) *n*-grams as stylistic features to investigate stylistic differences (if any) across the work. Furthermore, we address the distinction between narration and conversation, which has previously been omitted. We performed exploratory multivariate analyses and demonstrated that, despite internal differences, single-authorship is more likely from a statistical point of view. If Galba had contributed something to the last quarter of the work, it would have been minimal.

1 Introduction

Tirant lo Blanc, hereafter TLB, is a chivalry novel written in Catalan toward the end of the 15th century. Its first edition was printed in 1490 in Valencia, Spain, although it had been supposedly composed between 1460 and 1465 (Ferrando, 2012). The Medieval Catalan literary masterpiece was praised for its style as “the best book in the world” by the 17th-century Spanish writer Cervantes in chapter VI of *Don Quijote de la Mancha* (de Cervantes Saavedra, 1999). The work was deemed to be the very first modern novel in Europe by Peruvian Nobel laureate Mario Vargas Llosa, who rediscovered and acknowledged its literary values in recent times (Vargas Llosa, 2015).

Regarding its authorship, a sharp contradiction exists between the dedication at the beginning of the book and the colophon at its end, where information about the authorship and printing is provided. Joanot Martorell affirms in the dedication that he is solely responsible for the work (single-authorship hypothesis), whereas the colophon states that Martorell *translated* the first three of the four parts and that the fourth part was *translated* by Martí Joan de Galba (dual-authorship hypothesis). Here, *translation* refers to creation, as feigning a *translation* was commonplace during the period under consideration. The apparent inconsistency has been reconciled supposing that Martorell wrote most of the work and Galba revised and expanded it later (Martorell, 2008). Nonetheless, this supposition requires empirical validation to verify whether Galba actually participated in the creation and, if so, to identify Galba’s contributions.

Thus, this study revisits the unsettled authorship attribution of TLB by exploiting parts-of-speech (POS) *n*-grams as stylistic features. Existing literature has only considered a word-length distribution, that of the most frequently used words, and indices of the diversity of vocabulary. This study also addresses the distinction between narration and conversation, which has hitherto been disregarded, for fear that varying proportions of the two components in the work might confound the eventual outcome. We performed exploratory multivariate analyses and demonstrated that, despite internal differences, single authorship is more likely from a statistical point of view. If Galba had contributed something to the last quarter of the work, it would have been minimal.

The remainder of this paper is organized as follows. In Section 2, we review the existing literature and highlight its limitations. Section 3 describes the methodology followed in this study. In Section 4, we present the experimental results, followed by a discussion in Section 5. Finally, Section 6 concludes the paper.

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 16–26
December 1-3, 2023 ©2023 Association for Computational Linguistics

2 Related Work

The single-authorship hypothesis, according to which Martorell is the sole author, is based on the description in the dedication, whereas the dual-authorship hypothesis, according to which Martorell wrote the first three parts and Galba added the fourth, is indicated in the colophon. The single-authorship hypothesis has been endorsed by renowned philologist Martí de Riquer (de Riquer, 1990; Martorell, 2016), although the dual-authorship hypothesis has not been completely rejected (Martorell, 2008). Assuming that the dual-authorship hypothesis is true, the question arises as to where the fourth part that Galba purportedly composed begins. TLB is not explicitly divided into four parts, except for the first part, the beginning of which is noted ahead of chapter 1. Considering that TLB consists of 487 chapters of unequal length, de Riquer (1990) estimated that if the colophon is to be trusted, the fourth part begins with chapter 363 in terms of the number of chapters, or around chapter 283 in terms of the total length of the novel.

Under these circumstances, stylometry plays a key role (Martorell, 2008). Stylometry aims to identify the genuine author(s) of a written text through quantitative analysis (Stamatatos, 2009). A series of stylometric studies have delved into questions concerning the authorship of TLB. In their pioneering study, Girón et al. (2005) examined the distribution of word lengths (Mendenhall, 1887; Williams, 1975) and the most frequent context-free words, including articles, conjunctions, prepositions, and pronouns. They detected a change in the distribution of the variables from chapters 371 to 382 and concluded that the results corroborate dual authorship. Nonetheless, they admit the possibility that the observed differences may have been due to factors other than changes in authors.

One shortcoming of Girón et al. (2005) is that word-length distribution is not currently viewed as the most effective feature for authorship attribution tasks, which is implied by its practical absence in recent studies. Furthermore, the linguistic interpretation of word-length distribution is not straightforward. The stylistic information encoded therein is unclear.

Another drawback is the model selection process for the number of authors involved in the work. They compared two probabilistic models corresponding to the single and dual authorship hypotheses. The former consists of a single multinomial distribution, and the latter comprises a mixture of two distributions. Then, the ratio of posterior probabilities between the two models was computed to decide which one was more likely. However, they did not consider the model's complexity. Hence, the selection of the dual authorship hypothesis was the natural outcome, given that a more complex model fits better than a simpler one. The trade-off between model complexity and goodness of fit should be addressed appropriately.

In addition, the authors disregarded the distinction between narration and conversation when the analyses were vulnerable to the varying proportions of these two components in the work. In fact, the narration/conversation ratio fluctuates greatly among chapters, as depicted in Figure 1. The vertical axis represents the narration ratio, computed as the number of tokens in the narration divided by the chapter length. The curve represents the moving average with a window size of 20. The ratio of narration remains high from around chapter 375 onward to the end, whereas it is negligible from chapters 40–100. We assume that a different proportion of narration/conversation is not *per se* indicative of different authorship because its constancy across a work by a single author is not self-evident; narration/conversation may well be abundant in some sections and exiguous in others.

Using analogous approaches, other studies arrived at the same conclusion (Girón et al., 2005; Riba and Ginebra, 2005; Puig et al., 2015; Font et al., 2016). Riba and Ginebra (2006) also reinforces their findings using eight different indices of the diversity of vocabulary, which is rarely utilized as an effective stylistic feature either.

Thus, this study intends to shed new light on the authorship attribution of TLB in the following ways: (i) we leverage POS *n*-grams, which are effective and linguistically interpretable stylistic features; (ii) we conduct model selection appropriately, considering the trade-off between model complexity and goodness of fit; and (iii) we address the distinction between narration and conversation, which has hitherto been omitted.

3 Methods

A digitized transcription of the *princeps* edition was used in this study (Martorell, 2006)[1]. The

[1] `https://www.cervantesvirtual.com/obra/tirant-lo-blanc--1/`

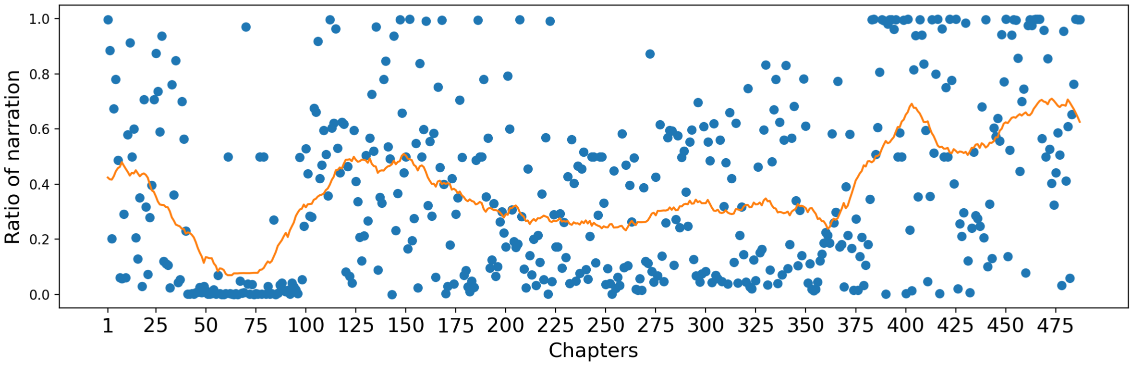


Figure 1: Ratio of narration along the chapters. The ratio was computed as the number of tokens in the narration divided by the chapter length. The curve represents the moving average with a window size of 20.

chapter titles and Latin phrases (e.g., *deo gracias* "thanks to God") were removed. We also eliminated paragraphs in the letter format that deviated from the typical structure of the work. Numerous passages allegedly plagiarized from other works (de Riquer, 1990) were retained as such for the sake of simplicity. Regarding punctuation, commas were eliminated so that editorial interventions would not come into play, whereas periods, colons, semi-colons, interrogation marks, and exclamation marks indicating sentence boundaries were retained as single punctuation symbols. Moreover, contracted and concatenated forms were separated prior to POS tagging (e.g., *l'art* "the art" and *donant-lo* "giving it" were divided into *l' art* and *donant -lo*, respectively).

Pre-processing resulted in 420,879 tokens and a vocabulary size of 17,181. For subsequent analyses, we did not adopt the original chapter division because the lengths varied considerably from one another. Instead, we generated equal-length pieces of 10K tokens to obtain reliable statistics. The length of 10K tokens is way above the minimum sample size of 5K tokens that was shown to be sufficient for stylometric analysis (Eder, 2015). The shortest final piece of 879 tokens was merged into the penultimate one. Thus, the entire work resulted in forty-two pieces.

We leveraged POS n-grams as stylistic features. The effectiveness of POS n-grams has been demonstrated by the previous research addressing literary works in multiple languages, including English (Koppel et al., 2002; Clement and Sharp, 2003; Juola, 2006; Hirst and Feiguina, 2007; Eder, 2015; Pokou et al., 2016; Savoy, 2017), French (Kocher and Savoy, 2019), Japanese (Uesaka and Murakami, 2015), and Spanish (Kawasaki, 2021, 2022). The advantages of employing POS sequences are multi-fold: (i) the numerous occurrences provide reliable statistics; (ii) they are relatively, if not completely, independent of content; (iii) they are deemed to be reliable style markers (Holmes, 1998; Juola, 2006; Stamatatos, 2009). Although partially, they capture syntactic patterns that are difficult to imitate and allegedly out of the conscious control of the author (Baayen et al., 1996); and (iv) they are supposedly less vulnerable to editorial interventions that would manipulate the original; in fact, orthographic vacillation could derive not only from the author but also from the typesetters in the Middle Ages.

The tokens were POS-tagged according to lemmatized concordance[2]. Specifically, we looked up each token in the concordance prepared in keywords in context format, considering its preceding and following contexts. Thus, more than 99% of the tokens were correctly tagged. Tokens that were ambiguous or left untagged were assigned a special tag, UNK, for simplicity, although manual tagging was desirable. Consequently, the number of POS tags amounted to thirteen[3]. For the most frequent twenty words, including adverbs, conjunctions, prepositions, and verbs, we adopted lemma forms in lieu of the POS-tags to exploit their particular usage[4]. For example, the preposition *de* "of"

[2] We are greatly indebted to Dr. Eduard Baile López of University of Alicante for providing us with the valuable data.

[3] ADJ(ECTIVE), ADV(ERB), ART(ICLE), CONJ(UNCTION), CONTR(ACTION BETWEEN PREPOSITION AND ARTICLE), INTERJ(ECTION), N(OUN), PREP(OSITION), PRON(OUN), PROPER(NOUN), PUNCT(UATION), UNK(NOWN), and V(ERB)

[4] *i* "and", *de* "of", *que* "that", *ésser* "to be", *en* "in", *a* "to", *per* "for", *no* "not", *fer* "to do", *haver* "to have", *tot* "all", *com* "as", *ab* "with", *dir* "to say", *molt* "much", *se* "oneself", *gran* "great", *un* "a", *qui* "who", and *tenir* "to have".

was not converted into PREP but maintained as such. This resulted in thirty-three unigram types in total: thirteen POS tags and twenty lemma forms.

For the subsequent multivariate analyses, every text piece was represented as a vector, with its elements being the z-transformed relative frequencies of the n-grams (Burrows, 2002). The relative frequencies were standardized to have a zero mean and unit variance for every variable. We considered only the most frequent POS n-grams above a given rank threshold r, whereas the remainder was aggregated under the OTHERS label. Thereafter, we performed two exploratory multivariate analyses, i.e., principal component analysis (PCA) and k-means clustering. As no other works by the relevant authors were available, it was infeasible to apply supervised methods such as classification. To assess the robustness of the analyses, we varied the n-gram size n for $n \in \{1, 2, 3, 4\}$ and the rank threshold r for $r \in \{50, 100, 300, 500\}$. For $n = 1$, r was fixed at 33, which is the number of unigram types.

4 Results and Analyses

In this section, we present the experimental results without a narration/conversation distinction. The results of the respective parts will be presented in Section 5.1. For illustrative purposes, we provide the results obtained with the hyper-parameters $(n, r) = (3, 300)$, unless noted otherwise.

First, we examined the overall similarity patterns across the entire work. Figure 2 displays the pair-wise distance scores between the pieces. The i-th piece is designated as TLB_i. The scores were calculated as $\sqrt{\|\boldsymbol{x_i} - \boldsymbol{x_j}\|^2/r}$, where $\boldsymbol{x_i}$ represents the feature vector for the i-th piece and r the number of n-grams considered. The bluer (redder) the cell, the more (less) similar the pair of pieces. We can readily discern a large cluster comprised of TLB_01–TLB_34, within which the distance scores are small compared to the rest of the pieces.

4.1 PCA

We performed PCA using `sklearn.decomposition.PCA` with the default settings (Pedregosa et al., 2011)[5]. Figure 3 illustrates the first two PC scores obtained with the hyper-parameters $(n, r) = (3, 300)$. The contribution ratios of PC1 and PC2 were 12.3% and 9.8%, respectively. Figure 3 apparently shows no significant pattern. However, we found that both PC1 and PC2 presented a moderate negative correlation with the proportion of conversational parts in the pieces: Spearman's $\rho = -0.66 \quad (p < 0.01)$ for PC1 and $\rho = -0.34 \quad (p = 0.03)$ for PC2. It is probable that the principal components simply reflect the proportion of narration/conversation, although it is not impossible that they reflect different authorship. Hence, we find it more practical to distinguish between the two parts and verify whether the same pattern emerges.

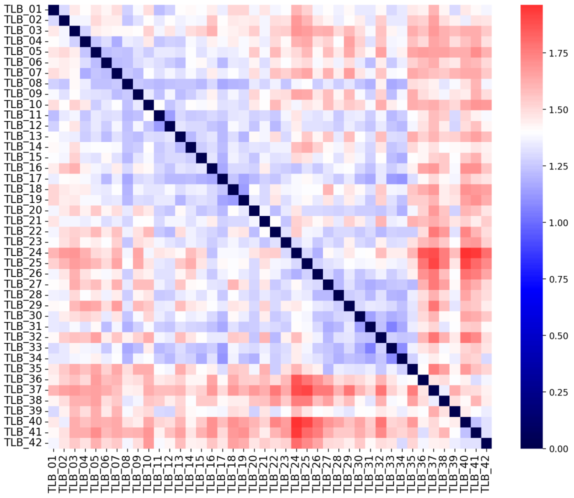


Figure 2: Pair-wise distance scores between the 10K-token pieces from the entire work, computed with hyper-parameters $(n, r) = (3, 300)$. The bluer (redder) the cell, the more (less) similar the pair of pieces.

4.2 k-means

We performed k-means clustering using `sklearn.cluster.KMeans` with the default settings (Pedregosa et al., 2011)[6]. The number of clusters k was fixed at $k = 2$, which is the supposed maximum number of authors involved. As the algorithm is sensitive to the initially selected centroids, we ran it 100 times to compute the mean concordance rate, which is defined as the average number of times a pair of pieces is found in the same cluster. Our premise was that no clear-cut pattern would emerge if stylistic differences did not exist.

Figure 4 presents the pair-wise mean concordance rates obtained with hyper-

[5] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[6] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

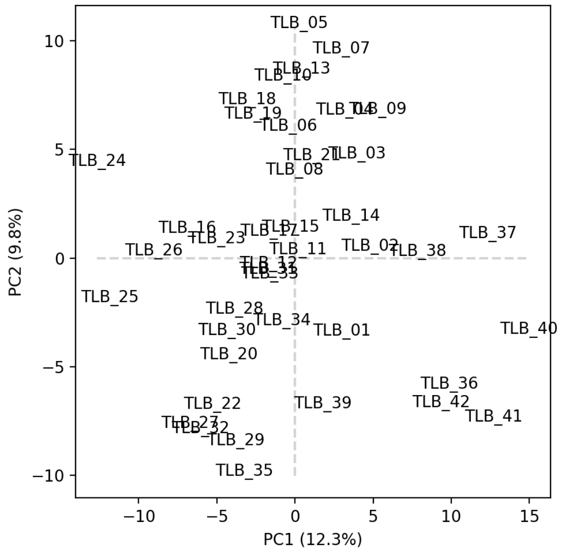


Figure 3: Scatter plot of PC1 and PC2 for the entire work with hyper-parameters $(n, r) = (3, 300)$.

parameters $(k, n) = (2, 3)$, while varying $r \in \{50, 100, 300, 500\}$. The darker the cell, the more often the pair of pieces were classified into the same cluster. Figure 4 illustrates that the clustering method is susceptible to the hyper-parameter r, resulting in inconsistent outcomes. The resulting clusters were also sensitive to n (data not shown). Consequently, it was difficult to draw definitive conclusions. If two distinct styles were to exist in the work, they would have been detected consistently regardless of different hyper-parameter settings, which was not the case.

For $r \in \{300, 500\}$, we can see a boundary between TLB_35 and TLB_36, which agrees with the findings of Riba and Ginebra (2005). They detected it in chapters 371–382, which roughly correspond to the second half of TLB_35 and the first half of TLB_36. However, this is also the point where the narration ratio increases (Figure 1). Therefore, we suspect that what Riba and Ginebra (2005) detected was not necessarily a change-point of authors but rather that of the narration/conversation ratio, and argue for the distinction between narration and conversation parts.

5 Discussion

5.1 Narration/Conversation Discrimination

As described above, the unequal amount of narration/conversation in the work potentially affects the resultant n-gram distribution. To avoid possible confounding effects, we distinguished between the narration and conversation sections. Identification of the two parts was readily made as the beginning of the conversation paragraphs is indicated with special characters. The entire text was first segregated into narration and conversation parts, and then each part was divided into 10K-token pieces. When the length of the last piece exceeded 6K, it was treated as an independent piece; otherwise, it was merged into the penultimate piece to prevent it from suffering data paucity. Thus, the narration and conversation parts resulted in eighteen and twenty-four pieces, respectively. The *i*-th piece in the narration (conversation) part was designated as TLB_N(C)_*i*.

5.1.1 Narration

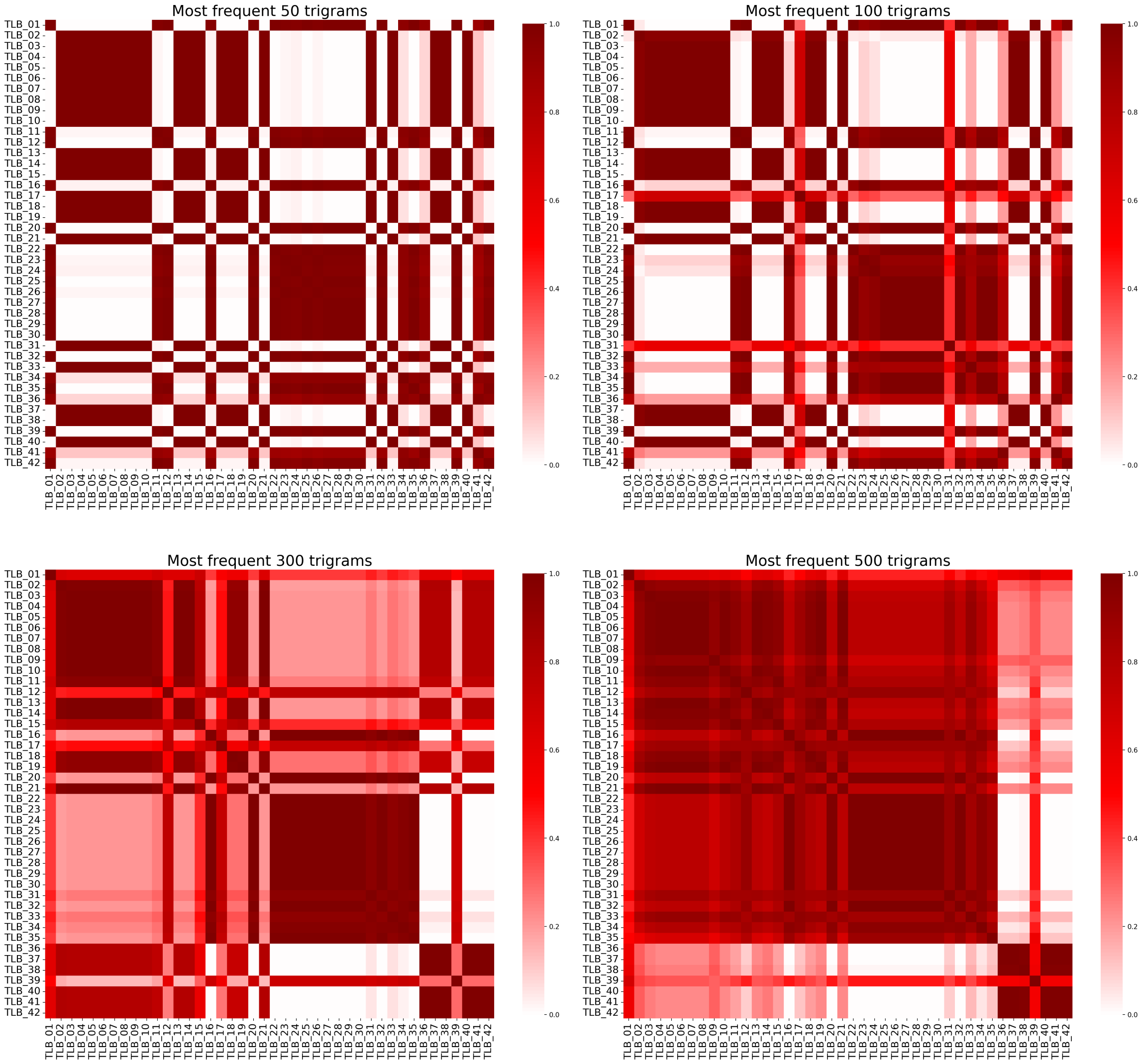
Figure 5 illustrates the first two PC scores for the narration part with the hyper-parameters $(n, r) = (3, 300)$. The contribution ratios of PC1 and PC2 were 18.0% and 10.2%, respectively. PC1 neatly separates TLB_N_14–TLB_N_18 on the far left side from the rest, whereas the interpretation of PC2 is difficult to make.

The pair-wise mean concordance rates among the narration parts are displayed in Figure 6. The results were relatively robust with other hyper-parameter settings. The narration section presents a clear boundary between TLB_N_13 and TLB_N_14, which approximately corresponds to chapter 350, where the story turns to the fate of *Plaerdemavida*. This boundary does not diverge greatly from the estimation by de Riquer (1990) that the beginning of the fourth part should be situated in chapter 363 in terms of the number of chapters. Furthermore, it accords with de Riquer's earlier opinion that Galba's contribution should be located from chapter 349 onward (Martorell and de Galba, 1947). In sum, the detected boundary does not contradict the description in the colophon that Galba created the fourth section.

5.1.2 Conversation

Figure 7 illustrates the first two PC scores for the conversation part with the hyper-parameters $(n, r) = (3, 300)$. The contribution ratios of PC1 and PC2 were 18.2% and 10.6%, respectively. PC1 separates TLB_C_02–TLB_C_06 on the far left side from the rest, and among which PC2 isolates TLB_C_22–TLB_C_24 from the remainder.

Next, we present the pair-wise mean concordance rates for the conversation component in Figure 8. The conversation part presents two bound-


Most frequent 50 trigrams

Most frequent 100 trigrams

Most frequent 300 trigrams

Most frequent 500 trigrams

Figure 4: Pair-wise mean concordance rates computed from 100 iterations of k-means. The hyper-parameters were set at $(k, n) = (2, 3)$ and $r \in \{50, 100, 300, 500\}$. The darker the cell, the more similar the pair of pieces.

aries: one between TLB_C_01 and TLB_C_02, which corresponds approximately to chapter 29, and the other between TLB_C_06 and TLB_C_07, which corresponds approximately to chapter 101. The results were relatively robust with other hyper-parameter settings.

The pieces TLB_C_02–TLB_C_06, or chapters 29–101, roughly correspond to the latter part of the section "William of Warwick" and the entire section of "Tirant in England" (de Riquer, 1990). These chapters are exceptional in that they consist of conversation only (Figure 1) and are characterized by an abundance of narrational components within conversation, in contrast to the dialogic style of the rest of the chapters. This peculiarity could be attributed to the alleged adaptation for TLB of *Guillem de Vàroich* (GV), which Martorell himself would have composed prior to the creation of TLB (Gili i Gaya, 1947; de Riquer, 1990)[7]. In such a case, the second boundary between TLB_C_06 and TLB_C_07 would not necessarily reflect different authorship but rather Martorell's internal stylistic variation.

Regarding the first boundary between TLB_C_01 and TLB_C_02, it is noticeable that TLB_C_01 corresponding to the first part of "William of Warwick" does not resemble its continuation but the rest of the work starting from TLB_C_07. We speculate that Martorell's intensive retouching of the aforementioned GV only involved its initial part to accommodate it to the newly composed TLB and that the rest was left relatively intact.

Notably, the narration and conversation diverged in terms of the boundary that separates the two in-

[7] http://www.cervantesvirtual.com/obra/guillem-de-varoich--0/

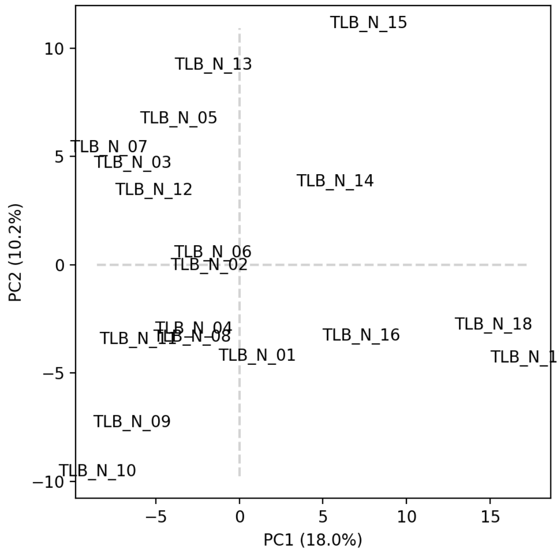


Figure 5: Scatter plot of PC1 and PC2 for the narration part with hyper-parameters $(n, r) = (3, 300)$.

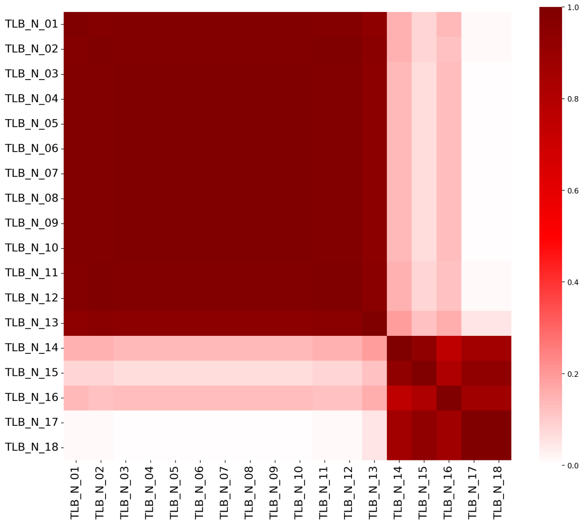


Figure 6: Pair-wise mean concordance rates for the narration part that was computed from 100 iterations of k-means performed with the hyper-parameters $(k, n, r) = (2, 3, 300)$. The darker the cell, the more similar the pair of pieces.

ternal clusters. Although this speculation requires verification by conducting experiments with undisputed works, we argue that if a different hand had come into play, both narration and conversation would coincide at the cluster boundary, which is not the case with $k = 2$. Nonetheless, when k is set to three for the conversation part, there emerges a subcluster within the second cluster, whereas the first cluster remains intact, as displayed in Figure 9. This subcluster comprises TLB_C_22–TLB_C_24, corresponding approximately to chapters 355–487. This boundary agrees well with that detected for the narration part at chapter 350, as noted above. In line with Martorell and de Galba (1947), the concurrence of the boundaries suggests that, if Galba had made some contribution to TLB, it should be located from chapter 350 to the end. The fact that new boundaries do not emerge when k is set to four or five points to strong internal cohesion of the clusters.

5.2 Detection of Number of Components

Thus far, it is evident that internal variation exists both in the narration and conversation parts. However, we are yet to verify if the variation is so large as to ascribe it to different authors. Despite the detection of the correct number of components being a challenging problem in stylometry (Koppel et al., 2011), we attempted to statistically determine the number of distinct hands that may have participated in TLB. We presumed that if single-authorship was more likely, only one component would be detected instead of two or more components, in which case multiple-authorship would be backed up.

By formulating the problem as model selection, we applied a Gaussian Mixture Model (GMM) combined with Bayes Information Criterion (BIC). GMM allows for probabilistic clustering to explore the heterogeneity in multivariate data (Frühwirth-Schnatter, 2006; Murphy, 2012). Combination with BIC enables model selection, considering the trade-off between model complexity and goodness of fit; a smaller BIC value indicates a better model. The capability of the algorithm to estimate the correct number of components has been demonstrated in the literature (Leroux, 1992). Although its effectiveness for stylometric studies requires empirical validation with the works of undisputed authorship, it will be worthwhile to apply the method to our case of interest. We implemented the algorithm using `sklearn.mixture.GaussianMixture` (Pedregosa et al., 2011)[8] with the default `full` covariance parameter and varying the number of components $k \in \{1, 2, 3, 4, 5\}$.

Figure 10 reveals that the effective number of components is $k = 1$ for every r in the narration. An identical pattern was observed for the conversation part. The behavior was consistent across the

[8] https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html

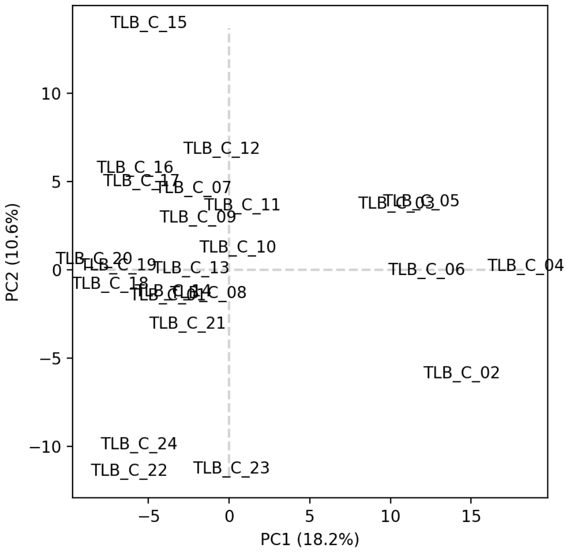


Figure 7: Scatter plot of PC1 and PC2 for the conversation part with hyper-parameters $(n, r) = (3, 300)$.

hyper-parameter space (n, r) (figures not shown) except for $(n, r) = (1, 33)$, in which case the estimated number of components was two for narration and three for conversation. The fact that the outcome converges as n grows larger would justify giving more importance to the results obtained with $n \geq 2$. We suspect that unigrams are too coarse-grained to elicit an immanent pattern.

Consequently, we argue that, despite internal differences, single-authorship is more likely than dual-authorship from a statistical viewpoint. We conjecture that the clear split observed with PCA and k-means simply reflects Martorell's internal stylistic variation without necessarily pointing to different authorship. Alternatively, Galba might have actually contributed to the creation of the fourth part starting from around chapter 350 onward to the end, but too little for his own stylistic fingerprints to be recognized.

5.3 Distinctive POS *n*-grams

We explored POS n-grams that played a crucial role in the multivariate analyses and deserve special mention from the philological viewpoint. As we explained in Section 3, the relative frequencies of n-grams were standardized to have zero mean and unit variance for every variable. An n-gram was considered distinctive when its absolute value was above 1 on average for the pieces of interest.

With respect to the narration part, we focus on the pieces TLB_N_14–TLB_N_18 forming a cluster in Figure 6. In these pieces, the trigrams that include ADJ_N are frequently used: MOLT_**ADJ_N**, **ADJ_N**_I, and **ADJ_N**_ADV. The sequence ADJ_N represents the preposition (instead of posposition) of an adjective to the noun that it modifies (e.g., *triümphal victòria* “triumphant victory”). Coromines (1971) attributed the excessive use of epithet preposition to the alleged Galba's contribution. Also characteristic are the sentences beginning with the conjunction *i* “and” followed by a verb, as illustrated by **PUNCT_I_V** and **PUNCT_I_FER** (e.g., . *E lexaren* “. And they left”). Other distinctive features include the use of the adverb *molt* “very”, as exemplified by **MOLT**_ADJ_V and **MOLT**_ADV_N (e.g., *molt bé acompanyats* “very well accompanied”), and that of the adjective *gran* “great”, as seen in **GRAN**_N_I (e.g., *gran importància e* “great importance and”).

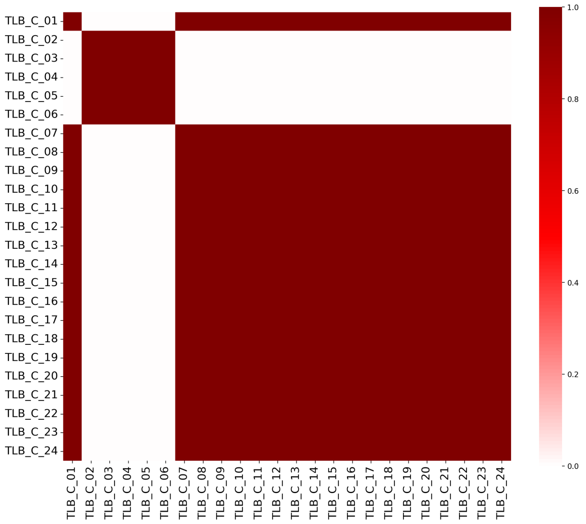


Figure 8: Pair-wise mean concordance rates for the conversation component that was computed from 100 iterations of k-means performed with the hyper-parameters $(k, n, r) = (2, 3, 300)$. The darker the cell, the more similar the pair of pieces.

In the conversation part, we first focus on the pieces TLB_C_02–TLB_C_06 forming a cluster in Figure 8. In these pieces, the trigrams that include ART_N representing a noun phrase headed by an article (e.g., *lo rey* “the king”) are extensively used: PUNCT_**ART_N**, **ART_N**_V, **ART_N**_I, COM_**ART_N**, **ART_N**_ÉSSER, ADV_**ART_N**, DE_**ART_N**, **ART_N**_ADV, AB_**ART_N**, and V_**ART_N**. This usage reflects the abundance of narrational components within the conversation. In contrast, the following trigrams appear much fewer times: PRON_**HAVER_V**, which contains present perfect construction formed by *haver* “to have” and

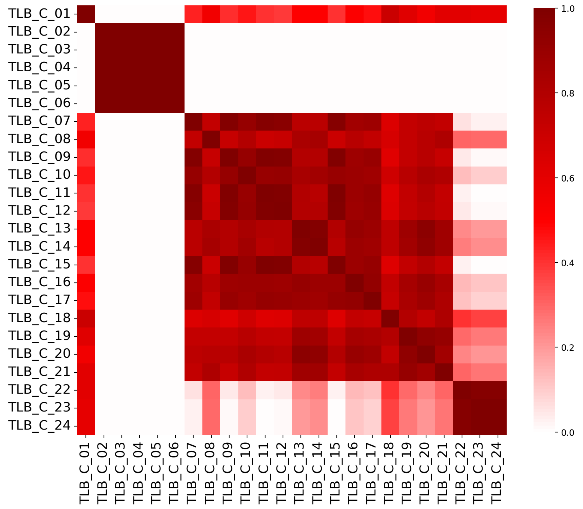


Figure 9: Pair-wise mean concordance rates for the conversation component that was computed from 100 iterations of k-means performed with the hyper-parameters $(k, n, r) = (3, 3, 300)$. The darker the cell, the more similar the pair of pieces.

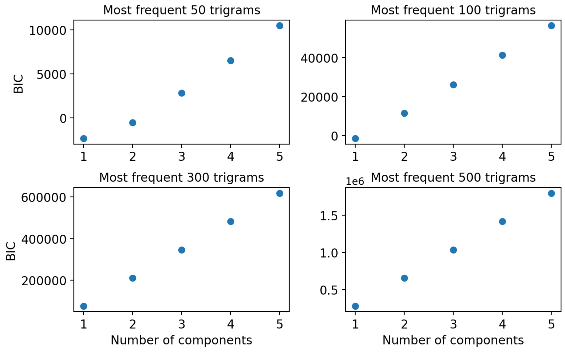


Figure 10: Model selection with Gaussian Mixture Model for narration part. The hyper-parameters are $n = 3$ and $r \in \{50, 100, 300, 500\}$. A smaller BIC value indicates a better model.

past participle (e.g., *·ns ha dats* “has given us”); V_**PUNCT_CONJ** and N_**PUNCT_CONJ**, which represent a sentence beginning with conjunction (e.g., *glòria. Donchs* “glory. So”); and **ADJ_N**_V, **ADJ_N**_CONJ, ART_**ADJ_N**, etc, all of which represent preposition of an adjective to the noun that it modifies.

The following trigrams characterize a subcluster TLB_C_22–TLB_C_24 in Figure 9 for frequent occurrences: **HAVER_V**_ART, which contains present perfect construction (e.g., *ha presa la* “has caught the”); **V_V_I**, which involves, for instance, an infinitive preceded by an auxiliary verb including *poder* “can” and *voler* “to want” (e.g., *podia veure e* “could see and”); and ART_**ADJ_N**, **ADJ_ADJ_N**, etc., all of which represent preposition of an adjective to the noun that it modifies. Its frequent use is also seen in the corresponding narration part (Coromines, 1971).

6 Conclusions and Future Work

This study revisited the unsettled authorship attribution of *Tirant lo Blanc* using stylometric techniques; specifically, we exploited POS n-grams as stylistic features. Furthermore, we addressed the distinction between narration and conversation, which has hitherto been omitted. We performed exploratory multivariate analyses and demonstrated that, despite internal differences, single-authorship is more likely from a statistical point of view. If Galba had contributed something to the last quarter of the work, it would have been minimal.

One limitation of our study is the adoption of rather coarse granularity in parts-of-speech. For instance, we did not distinguish between verbal forms such as finite forms, infinitive, gerund, and participle and instead treated them all under the category of VERB. However, their peculiar usage has been pointed out in previous literature (Gili i Gaya, 1947; Ferrando, 1987; de Riquer, 1990; Ferrando, 2012) and so could be useful for detecting authorial fingerprints as well. Furthermore, it will be intriguing to explore the orthographic, lexical, morphological, and syntactic traits that have been suggested as distinctive in previous research (Gili i Gaya, 1947; Coromines, 1971; Ferrando, 1987; Skubic, 1989; de Riquer, 1990; Ferrando, 2012), to name a few[9].

Moreover, two hypotheses concerning the genesis of TLB remain to be examined stylometrically: (i) that a short fragmentary manuscript denominated *Guillem de Vàroich* was actually written by Martorell (Gili i Gaya, 1947; de Riquer, 1990); and (ii) that the Valencian writer Joan Roís de Corella is the genuine author of TLB (Guia i Marín, 1996).

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP18K12361 and JP23K12152.

References

Harald Baayen, Hans van Halteren, and Fiona Tweedie. 1996. Outside the Cave of Shadows: Using Syntac-

[9] See also: http://www.cervantesvirtual.com/portales/joanot_martorell_i_el_tirant_lo_blanc/llengua/

tic Annotation to Enhance Authorship Attribution. *Literary and Linguistic Computing*, 11(3):121–132.

John Burrows. 2002. 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17(3):267–287.

Miguel de Cervantes Saavedra. 1999. El Ingenioso Hidalgo Don Quijote de la Mancha.

Ross Clement and David Sharp. 2003. Ngram and Bayesian Classification of Documents for Topic and Authorship. *Literary and Linguistic Computing*, 18(4):423–447.

Joan Coromines. 1971. Sobre l'estil i manera de Martí J. de Galba i el de Joanot Martorell. In *Lleures i converses d'un filòleg*, pages 363–378. Club Editor, Barcelona.

Maciej Eder. 2015. Does Size Matter? Authorship Attribution, Small Samples, Big Problem. *Digital Scholarship in the Humanities*, 30(2):167–182.

Antoni Ferrando. 1987. Entorn de la llengua del *Tirant lo Blanc*. *Estudis Romànics*, 4:369–372.

Antoni Ferrando. 2012. Llengua i context cultural al *Tirant lo Blanc*. Sobre la identitat del darrer Joanot Martorell (1458-1465). *eHumanista*, 22:623–668.

Marti Font, Xavier Puig, and Josep Ginebra. 2016. Bayesian Analysis of the Heterogeneity of Literary Style. *Revista Colombiana de Estadística*, 39(2):205–227.

Sylvia Frühwirth-Schnatter. 2006. *Finite Mixture and Markov Switching Models*. Springer, New York, NY.

Samuel Gili i Gaya. 1947. Noves recerques sobre *Tirant lo Blanch*. *Estudis Romànics*, 1:135–147.

Javier Girón, Josep Ginebra, and Alex Riba. 2005. Bayesian Analysis of a Multinomial Sequence and Homogeneity of Literary Style. *American Statistician*, 59(1).

Josep Guia i Marín. 1996. *De Martorell a Corella. Descobrint l'autor del Tirant lo Blanc*. Editorial Afers, Barcelona.

Graeme Hirst and Ol'ga Feiguina. 2007. Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts. *Literary and Linguistic Computing*, 22(4):405–417.

David I. Holmes. 1998. The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing*, 13(3):111–117.

Patrick Juola. 2006. Authorship Attribution. *Foundations and Trends in Information Retrieval*, 1(3):233–334.

Yoshifumi Kawasaki. 2021. Stylometric Analysis of Avellaneda's *Don Quijote*. In *12th International Conference on Corpus Linguistics*, Universidad de Murcia (Online). Spanish Association for Corpus Linguistics.

Yoshifumi Kawasaki. 2022. A Stylometric Analysis of *Amadís de Gaula* and *Sergas de Esplandián*. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 1–7. Association for Computational Linguistics.

Mirco Kocher and Jacques Savoy. 2019. Evaluation of Text Representation Schemes and Distance Measures for Authorship Linking. *Digital Scholarship in the Humanities*, 34(1):189–207.

Moshe Koppel, Navot Akiva, Idan Dershowitz, and Nachum Dershowitz. 2011. Unsupervised decomposition of a document into authorial components. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1356–1364, Portland, Oregon, USA. Association for Computational Linguistics.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, 17(4):401–412.

Brian. G. Leroux. 1992. Consistent Estimation of a Mixing Distribution. *The Annals of Statistics*, pages 1350–1360.

Joanot Martorell. 2006. *Tirant lo Blanc*. Alacant : Biblioteca Virtual Joan Lluís Vives, 2006.

Joanot Martorell. 2008. *Tirant lo Blanch*. Tirant lo Blanch, València.

Joanot Martorell. 2016. *Tirant lo Blanc*. Labutxaca, Barcelona.

Joanot Martorell and Martí Joan de Galba. 1947. *Tirant lo Blanc: Text, introducció, notes i índexs*. Editorial Selecta, Barcelona.

T. C. Mendenhall. 1887. The Characteristic Curves of Composition. *Science*, 9(214S):237–246.

Kevin P Murphy. 2012. *Machine Learning: A Probabilistic Perspective*. The MIT Press, Cambridge, MA.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Yao Jean Marc Pokou, Philippe Fournier-Viger, and Chadia Moghrabi. 2016. Authorship Attribution Using Variable Length Part-of-Speech Patterns. In *Proceedings of the 8th International Conference on Agents and Artificial Intelligence*, volume 2, pages 354–361.

Xavier Puig, Martí Font, and Josep Ginebra. 2015. Classification of Literary Style that takes Order into Consideration. *Journal of Quantitative Linguistics*, 22(3).

Alex Riba and Josep Ginebra. 2005. Change-point estimation in a multinomial sequence and homogeneity of literary style. *Journal of Applied Statistics*, 32(1).

Alex Riba and Josep Ginebra. 2006. Diversity of Vocabulary and Homogeneity of Literary Style. *Journal of Applied Statistics*, 33(7).

Martí de Riquer. 1990. *Aproximació al Tirant Lo Blanc*. Quaderns Crema, Barcelona.

Jacques Savoy. 2017. Analysis of the Style and the Rhetoric of the American Presidents over Two Centuries. *Glottometrics*, 38:55–76.

Mitja Skubic. 1989. L'estructuració de l'oració composta en el *Tirant lo Blanc*. *Linguistica*, 29(1):137–145.

Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Ayaka Uesaka and Masakatsu Murakami. 2015. Verifying the authorship of Saikaku Ihara's work in early modern Japanese literature; A quantitative approach. *Digital Scholarship in the Humanities*, 30(4):599–607.

Mario Vargas Llosa. 2015. *Carta de batalla por Tirant lo Blanc*. Debolsillo.

C. B. Williams. 1975. Mendenhall's Studies of Word-Length Distribution in the Works of Shakespeare and Bacon. *Biometrika*, 62(1):207–212.

Translation from Historical to Contemporary Japanese Using Japanese T5

Hisao Usui and **Kanako Komiya**
Tokyo University of Agriculture and Technology
h-usui@st.go.tuat.ac.jp
kkomiya@go.tuat.ac.jp

Abstract

This paper presents machine translation from historical Japanese to contemporary Japanese using a Text-to-Text Transfer Transformer (T5). The result of the previous study that used neural machine translation (NMT), Long Short Term Memory (LSTM), could not outperform that of the work that used statistical machine translation (SMT). Because an NMT model tends to require more training data than an SMT model, the lack of parallel data of historical and contemporary Japanese could be the reason. Therefore, we used Japanese T5, a kind of large language model to compensate for the lack of data. Our experiments show that the translation with T5 is slightly lower than SMT. In addition, we added the title of the literature book from which the example sentence was extracted at the beginning of the input. Japanese historical corpus consists of a variety of texts ranging in periods when the texts were written and the writing styles. Therefore, we expected that the title gives information about the period and style, to the translation model. Additional experiments revealed that, with title information, the translation from historical Japanese to contemporary Japanese with T5 surpassed that with SMT.

1 Introduction

This paper develops a translation system from historical Japanese to contemporary Japanese. In Japan, there are a large number of historical literature including untranslated ones. Although the experts are able to translate them, it is not only difficult but also time-consuming to translate them manually. Therefore, translation systems of historical Japanese are necessary.

To our knowledge, two works have been done for translation from historical Japanese to contemporary Japanese (see Section 2). Hoshino et al. (2014) used statistical machine translation (SMT), which presented a translation system whose BLEU score (Papineni et al., 2002) was 28.02, the highest in the previous study. Takaku et al. (2020) used neural machine translation (NMT), Long Short Term Memory (LSTM), and the BLEU score of their system was 19.95. We believe that the lack of translation data is why the NMT, specifically, translation with LSTM, could not outperform the SMT. An NMT model tends to require more training data than an SMT model does but the parallel data of historical and contemporary Japanese are limited; they are 86,684 sentence pairs. Therefore, we used Text-to-Text Transfer Transformer (T5) (Raffel et al., 2020), which is a pre-trained large language model (LLM), to compensate for the lack of data. After the release of the GPT (Radford et al., 2018) and the BERT (Devlin et al., 2019) in 2018, LLMs have achieved state-of-the-art results in various tasks of natural language processing. We hypothesise that the poor performance due to the limited training data could be complemented by the use of large pre-trained models (see Section 3).

In addition, we focused on the diversity of the historical texts. The Japanese historical corpus consists of a variety of texts ranging in periods when the texts were written and the writing styles. For example, The Complete Collection of Japanese Classical Literature published by Shogakukan(SHOGAKUKAN, 2010) [1] contains older texts written in the 800s and relatively new texts written in the 1800s. In addition, the collection consists of texts of various styles, including diaries, poems, folk stories, and so on (see Section 4). Therefore, it is not surprising if the diversity of the texts gives a negative influence on the performance of the translation system.

Li et al. (2016) reported that giving speakers' IDs to persona-based dialogue generation improves the dialogue quality when the training data consists of data from multiple speakers. We believe that this is because the information of the speakers alleviates the negative influence of the diversity of the train-

[1] https://japanknowledge.com/en/contents/koten/

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 27–35
December 1-3, 2023 ©2023 Association for Computational Linguistics

ing data. Instead of speakers' IDs, we utilized the titles of the literature book from which the example sentences were extracted.

We compared our methods with the previous studies using BLEU and BERT scores and analysed the quality of the translations (see Sections 5, 6 and 7).

The contributions of this paper are as follows:

1. We developed a translation system from historical Japanese to contemporary Japanese using T5;
2. We proposed giving the title information of the literature book from which the example sentence was extracted to the translation model;
3. The BLEU score of our system outperformed those of the previous research; and
4. We discussed the translation quality using some examples.

2 Related Work

Research on machine translation from historical Japanese into contemporary Japanese has been reported by Hoshino et al. (2014) who used SMT and Takaku et al. (2020) who used LSTM. The BLEU scores were 28.02 for SMT and 19.95 for LSTM. We used T5 to complement the lack of parallel data that is required for the training of the model.

An example of a previous study of translation using T5 is work reported by Emezue and Dossou (2021). They used multilingual T5 (mt5) to translate African languages such as Kosa, Yoruba, and Igbo into English and French. The number of data on these African languages is limited; Kosa, the largest corpus has 158,660 monolingual sentences and 137,000 parallel data with English. They reported that the BLEU score of the Kosa-English translation was 30.25. Their research has shown that T5 is effective for low-resource language translation. Agarwal et al. (2020) used T5 for machine translation to aid bilingual data-to-text generation and semantic parsing. They showed the machine translation using T5 improved performances of generation and parsing tasks.

We gave the title of the literature book from which the example sentence was extracted to the translation model. This is inspired by Li et al. (2016), who gave speakers' IDs to the persona-based dialogue generation system. The dialogue quality improved when the training data consisted of data from multiple speakers; the utterances output from the system became more speaker-specific. We expected that the information of the speakers would alleviate the negative influence of the diversity of the training data. Instead of speakers' IDs, we input the titles of the literature book to alleviate the negative influence of the diversity of the training sentences in period and style. In addition, Caswell et al. (2019) added an extra token to back-translation data for noising techniques [2]. They inform the model which data is back-translated to alleviate the noise's effect. Instead, we added the title of the literature book to inform the model the period and writing style information.

In addition, there is research on translation related to digital humanities such as (Gupta, 2022), (Zheng et al., 2022), and (Piper and Erlin, 2022).

3 Translation Using T5

T5 is a pre-trained model based on Transformer (Vaswani et al., 2017) trained with Colossal Clean Crawled Corpus (C4), which is a cleaned version of Common Crawl's web crawl corpus. Using such high-quality data and careful tuning of the model, T5 achieved state-of-the-art performance on 26 various tasks in 2019.

In this paper, we fine-tuned pre-trained Japanese T5 to translate historical Japanese into contemporary Japanese. We aim to compensate for the lack of data using a pre-trained LLM because the amount of parallel corpus of historical and contemporary Japanese is limited. We used sonoisa/t5-base-japanese[3] for Japanese T5. This model was trained with Japanese Wikipedia dump data, Japanese OSCAR corpus, and Japanese CC-100: Monolingual Datasets from Web Crawl Data. Although it is mostly trained with contemporary Japanese, we used it for transformation from historical Japanese to contemporary Japanese. Therefore, this method could be deemed as a diachronic domain adaptation. We employed this method because Komiya et al. (2022) reported that the performance of Japanese word sense disambiguation of historical Japanese significantly improved when they used diachronic domain adaptation using Japanese BERT trained with contemporary

[2] We attempted the back translation for this research but it did not work. We think that this is because the data we used, Aozora-bunko, is not enough similar to the original corpus of historical Japanese

[3] https://huggingface.co/sonoisa/t5-base-japanese

Parallel data	
Historical sentence Contemporary sentence	而るに、既に講の終る日に成て、道俗男女員不知ず参り集たり。 さて、いよいよ講の終る日になると、僧俗・男女を問わず、 数知れぬほどの人々が参詣してきた。
Title-added parallel data	
Historical sentence Contemporary sentence	今昔物語集**(Konjaku)02:**而るに、既に講の終る日に成て、道俗男女員不知ず参り集たり。 さて、いよいよ講の終る日になると、僧俗・男女を問わず、 数知れぬほどの人々が参詣してきた。

Table 1: Example of parallel data and title-added data. The only difference between the parallel data and the title-added data is the prefix of the historical sentence: the title and colon(:).

Historical Japanese	
Total Number of Sentences	86,684
Vocabulary Size	49,200
Number of Tokens	2,774,745
Contemporary Japanese	
Total Number of Sentences	86,684
Vocabulary Size	45,690
Number of Tokens	3,611,783

Table 2: Parallel corpus of historical and contemporary Japanese. The data for translation are extracted from the Complete Collection of Japanese Classical Literature.

Japanese.

In addition, we gave the titles of the literature books from which the example sentences were extracted, to the translation model: the fine-tuned T5. We set a title and colon as a prefix of the historical Japanese sentence. Table 1 shows an example of the input: the prefix and text. The titles give information on the period and writing style of the original sentence to the model. The meanings of words sometimes vary depending on the writing style of the texts and the period when the texts were written. Therefore, we believe that this kind of information is useful for translation. It helps the model appropriately translate texts from various literature books written in a wide range of periods and in various writing styles. We were inspired by Li et al. (2016), who gave speakers' IDs to the persona-based dialogue generation system.

4 Data

We used a parallel corpus extracted from The Complete Collection of Japanese Classical Literature published by Shogakukan by Hoshino et al. (2014). This corpus consists of historical Japanese sentences paired with manually translated contemporary Japanese sentences. Table 2 shows the statistics of the parallel data of historical and contemporary Japanese. This is a diachronic corpus consisting of works from various periods.

We used 50 pieces of literature listed in Tables 3 and 4. However, they are 60 books because some literature has more than one volume and Hōgen monogatari and Heiji monogatari were compiled together in one book [4].

The corpus contains a total of 86,684 sentence pairs, some of which have multiple contemporary Japanese translations for a single historical Japanese. In the corpus, 221 historical Japanese data and 163 contemporary Japanese data were duplicated. Notably, these different translations of the same historical Japanese could decrease the BLEU score when they are split into the training and test data. The data was split into (training: development: test) = (82,591: 2,000: 2,093) sentence pairs, following (Takaku et al., 2020) To avoid data bias, the data were randomly split.

We made another data where the titles of this sentence were added to the head of historical Japanese from The Complete Collection of Japanese Classical Literature. We added the title and the mark colon (:) to the head of every historical sentence. Table 1 shows examples from both data. We appended the title to the beginning of each sentence pair. The title of the literature book is only added to the beginning of the historical Japanese and not to the contemporary Japanese.

As the parallel corpus did not have the title information, we searched the sentences of the source language, i.e., historical Japanese, in The Complete Collection of Japanese Classical Literature, the original dataset, and automatically identified it. In this process, a few errors happened especially for

[4] Chikamatsu-Monzaemon-shū has three volumes, Genji monogatari has six volumes, Heike monogatari has two volumes, and Kokin Wakashū has two volumes.

Titles	Year	Style
Chikamatsu-Monzaemon-shū (Collection of Chikamatsu-Monzaemon's Stories)	1703-1720	Playbook
Genji monogatari (The Tale of Genji)	Before 1001	Tale
Gikeiki (The Chronicle of Yoshitsune)	Before 1400	War chronicle
Heichū Monogatari (Tales of Heichū)	Around 1000	Poem tales
Heiji monogatari (The Tale of Heiji)	Mid-13th century	War chronicle
Heike monogatari (The Tale of the Heike)	Before 1309	Epic account
Hōgen monogatari (The Tale of Hōgen)	Around 1220	War chronicle
Hōjōki (Square-jō record)	1212	Essay
Imose Yamaonna Teikin	1771	Playbook
Ise monogatari (The Tales of Ise)	Around 900?	Poem tales
Izumi Shikibu Nikki (The Diary of Lady Izumi)	1008	Poetic diary
Jikkinsyo	1252?	Folktales
Kagerō Nikki (The Mayfly Diary)	974	Poetic diary
Kaikou	1764	Playbook
Kanameishi	1663	Folktales
Kindaishōka	1209	Essay on poetry
Kokin Wakashū (Collection of Japanese Poems Of Ancient and Modern Times)	905	Anthology of the poetry
Kokin Wakashū Appendix	905	Anthology of the poetry
Kokka-Hachiron	1742	Essay on poetry
Konjaku Monogatarishū (Anthology of Tales from the Past)	Around 1120	Folktales
Koraifūteishō	1197	Essay on poetry
Maigetsu-shō	1219	Essay on poetry
Makura no Sōshi (The Pillow Book)	1001	Essay
Murasaki Shikibu Nikki (The Diary of Lady Murasaki)	1010	Diary
Matsuranomiya Monogatari (The Tale of the Matsura Palace)	1201	Tale
Mutsuwaki (Chronicle of Mutsu)	Late 11th century	War chronicle
Nihon Ryōiki (Miraculous Stories from The Japanese Buddhist Tradition)	822	Folktales
Nii-Manabi-Iken	1811	Essay on poetry
Ochikubo Monogatari (The Tale of Ochikubo)	Around 1000	Tale
Ōkagami (The Great Mirror)	1119	History book
Otogi Monogatari	1678	Folktales
Sanukinosuke Nikki (The Diary of Sanukinosuke)	1109	Diary
Sarashina Nikki (The Sarashina Diary)	1020	Nonfiction narrative
Shasekishū (Sand and Pebbles)	1283	Buddhist text
Shōbōgenzō Zuimonki (The Treasury of the True Dharma Eye: Record of Things Heard)	1235-1238	Dharma talks
Shōmonki, Masakadoki (Chronicle of Masakado)	Before 1099	War chronicle
Soga Monogatari (The Revenge of the Soga brothers)	Before 1285	War chronicle
Taketori Monogatari (The Tale of the Bamboo Cutter)	Around 900	Fictional prose narrative
Takitsuke Moekui Keshizumi	1677	Folktales

Table 3: List of the Literature Books 1

Titles	Year	Style
Tannishō (Lamentations of Divergences)	Around 1300	Buddhist text
Tosa Nikki (Tosa Diary)	934	Poetic diary
Toshiyori Zuinō	1113	Essay on poetry
Tsurezuregusa (Essays in Idleness)	1332	Essay
Tsutsumi Chūnagon Monogatari (Tales of the Riverside Middle Counselor)	Before 1271	Short stories
Tsutsumi Chūnagon Monogatari Appendix	Before 1271	Short stories
Uji Shūi Monogatari	1221	Short stories
Ukiyo Monogatari	1661	Folktales
Utsubo Monogatari (The Tale of the Hollow Tree)	Around 1000	Tale
Yamato Monogatari (Tales of Yamato)	9th - 10th century	Short stories
Yōkyoku-shū	Before 1573	Playbook

Table 4: List of the Literature Books 2

short sentences, because sometimes the same short sentences appeared in multiple pieces of literature. However, we used them without any corrections.

5 Experiments

We conducted two kinds of experiments: (1) the experiment using the parallel data and (2) the experiment using the title-added parallel data.

We used sonoisa/t5-base-japanese and T5 tokenizer in the hugging face library for Japanese T5 and its tokenizer. We conducted a grid search for the learning rate, epoch number, and repetition penalty in both experiments as shown in Table 5. We set the hyperparameters of T5 as shown in Table 6. The other parameters are set to the default values. Because sonoisa/t5-base-japanese model does not need lots of memory, we trained the models by using only an RTX3090. It took 2.5 hours per epoch.

We measured BLEU scores and the similarities between translations using Sentence-BERT (Reimers and Gurevych, 2019) for evaluation. For the calculation of BLEU score, we used Sacreblue (Post, 2018) [5]. We calculated the similarities between translations for reference because BLEU scores do not directly reflect the meanings of the sentences. We used sonoisa/sentence-bert-base-ja-mean-tokens-v2 to calculate the similarities[6]. We used default settings for the hyperparameters of the Sentence-BERT.

6 Results

6.1 BLEU scores

Table 7 shows the best BLEU scores of the experiments using the parallel and title-added parallel data and their hyperparameters. It shows that the BLEU score of NMT with simple T5 using only the parallel data (27.50) is better than that of LSTM (19.95) (Takaku et al., 2020) and slightly lower than that of SMT (28.02) (Hoshino et al., 2014). It also shows that the BLEU score is better when the title-added parallel data was used than when the parallel data without the title was used. The BLEU score of NMT with T5 using the title-added parallel data (28.67) outperformed those of LSTM and SMT.

6.2 Similarities Using Sentence-BERT

Table 8 shows the average similarities between the translation of the systems and the reference translation calculated using Sentence-BERT. They were averaged over all the test examples. The BLEU scores are also shown in the table as a reference.

As shown in Table 8, the average similarity between the translation by the system trained with the parallel data and the reference translation was 0.787 and that between the translation by the system trained with the title-added parallel data and the reference translation was 0.784. The difference was 0.004 points, which is not significant. However, the system trained with the title-added parallel data outperformed the system trained with parallel data again.

[5] https://huggingface.co/spaces/evaluate-metric/sacrebleu

[6] https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2

Learning rate	0.0001, 0.0002, 0.0003, 0.0004, 0.0005, 0.0007, 0.001
Epoch number	1, 5, 10
Repetition penalty	1, 1.5, 2.0

Table 5: The hyperparameters of the experiments

max_seq_length	512
weight_decay	0.0
adam_epsilon	1e-8
warmup_steps	100
batch_size	128
gradient_accumulation_steps	4
n_gpu	1
early_stop_callback	False
fp_16	False
opt_level	01
max_grad_norm	1.0

Table 6: Hyperparameters of T5

7 Translation Analysis

Table 9 shows the translation examples of the experiments using the parallel and title-added parallel data. The first example is from Heiji Monogatari (The Tale of Heiji) and the second example is from Genji Monogatari (The Tale of Genji) [7]. In these two examples, we can see the reference translations are idiomatic, rather than literal translation.

For instance, in the first example, the original sentence does not contain the subject, who brought Tokiha (a woman) to someone who ordered someone to do so. Because omission of the subject often occurs in Japanese, this is a natural Japanese sentence. Nevertheless, the reference translation includes the subject; it says that Ito Musha, a specific warrior, brought her to someone who ordered him to do so. The translations of our systems do not include the subject, like the original sentence. In addition, the verb in the original sentence, "宣ふ", the basis form of "宣へ," literally means "say" rather than "order." However, the reference translation uses the verb "order" instead of "say" in this example, whereas the translations of the systems use the word "say." We think that this is probably because "order" is easier to understand than "say" in this context. However, it is an idiomatic translation again.

In addition, in the second example, the original sentence says "ここなる人々" literally means "people staying here." The translations of our systems are correct if the sentence was translated literally. However, the reference translation rephrased the word to "my sons." This is an additional explanation by a translator. Therefore, we believe that the BLEU scores of our systems tend to decrease because the reference translations are idiomatic. To evaluate these systems, literal translations are necessary.

Moreover, some Japanese verbs totally change the word when an honorific form is used. It is like the past tense of irregular verbs in English. In addition, some Japanese verbs have some variant honorific forms.

In the first example, two verbs, which mean bring someone and order, are written in normal form in the reference translation. However, when the T5 translation with the parallel data was used for the experiments, two verbs, which means bring someone and say, were written in honorific form. When the title-added parallel data was used, a verb that means "say" was written in another honorific form and the other two verbs, which mean bring someone and follow, were written in normal form.

In the second example, "おはせましかば," which means "were still alive" with old honorific form in the original sentence was translated into "生きていらっしゃったら," in the reference translation, which means the same but with contemporary honorific form. However, it was translated into "ご存命でいらっしゃったら" using our systems instead of "生きていらっしゃったら." "ご存命でいらっしゃったら" means "were still alive" again, and this is another expression that means the same with a different contemporary honorific form. In other words, these two expressions mean the same, however, the ways of honorification are different. These kinds of variations of expressions also tend to decrease the BLEU scores. Because we do not have the translation examples of the other systems, we cannot compare the results directly [8]. However, we believe that NMT tends to

[7] The title we used was Genji05, because we had 6 volumes for Genji Monogatari and it was the fifth one.

[8] We could not find any translation examples as same as (Takaku et al., 2020) because the dataset was the same but the

	Learning rate	Epoch number	Repetition penalty	BLEU
The parallel data	0.0002	10	1.0	27.50
The title-added parallel data	0.0005	5	1.5	28.67
LSTM (Takaku et al., 2020)				19.95
SMT (Hoshino et al., 2014)				28.02

Table 7: The best results of experiments using the parallel and the title-added parallel data

Data	Sentence-BERT	BLEU
Parallel data	0.784	27.49
Title-added data	0.787	28.57

Table 8: The average similarities of the translation of systems and the reference translation calculated using sentence-BERT

translate the texts with more variations than SMT does because SMT uses a dictionary. Therefore, these kinds of variations of expression probably happened more when NMT was used.

8 Conclusion and Future Work

In this paper, we translated historical Japanese into contemporary Japanese by fine-tuning Japanese T5, which is a large-scale pre-trained model. We expected that T5 would be useful for the translation because it compensates for the lack of parallel data. In addition, we proposed to add a book title from which the source sentence was extracted at the beginning of the input sentences. Because CHJ, the historical text corpus we used, ranges in periods when the texts were written and styles, we expected that the title of the books could give those kinds of information to the translation systems. The experiments revealed that T5 and giving the title of the books are effective for the translation; BLEU scores and the Sentence-BERT similarities of our system outperformed those of previous studies, which used SMT and LSTM, respectively. According to the analysis of the translation examples, we observed that the translations of our systems are more literal than the reference translations. However, the BLEU scores tend to decrease because the reference translations are idiomatic, rather than literal translation. The evaluation using the literal translations by human annotators is our future work.

In addition, further experiments and analyses should be done to investigate the effectiveness of our method. For example, comparison to the setting where random titles were used and the setting where only periods and genres are given should be explored. Moreover, we are planning to compare our method to Large Language Models, such as chatGPT.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 22K12145, and collaborative research projects of National Institute for Japanese Language and Linguistics. We would like to thank Hikaru Yokono, who gave us helpful advice and the data set used in previous research.

References

Oshin Agarwal, Mihir Kale, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2020. Machine translation aided bilingual data-to-text generation and semantic parsing. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 125–130, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Chris Chinenye Emezue and Bonaventure F. P. Dossou. 2021. MMTAfrica: Multilingual machine translation for African languages. In *Proceedings of the Sixth Conference on Machine Translation*, pages 398–411, Online. Association for Computational Linguistics.

Kshitij Gupta. 2022. MALM: Mixing augmented language modeling for zero-shot machine translation. In *Proceedings of the 2nd International Workshop*

data split was different. (Hoshino et al., 2014) did not contain the translation examples.

	Sentence	BLEU	SBERT
Original English Translation	具して参れ」と宣へば、常葉を具して参りけり。 (Someone,the subject in omitted here) brought Tokiha (This is a woman's name.) to him because (he, the subject is omitted here.) said "Bring her to me."		
Reference English Translation	連れて来い」と命ずるので、伊藤武者は常葉を連れてきた。 Ito Musha (This is a warrior's name.) brought Tokiha to him because (he, the subject is omitted here.) ordered as "Bring her to me."		
The parallel data English Translation	連れて参れ」と言われるので、常葉を連れて参った。 (He, the subject is omitted here.) brought Tokiha to him because (he, the subject is omitted here again.) said "Bring her to me."	21.98	0.81
The title-added parallel data English Translation	ついて来い」とおっしゃるので、常葉を連れて行った。 (He, the subject is omitted here.) brought Tokiha to him because (he, the subject is omitted here again.) said "Follow me."	32.04	0.81

	Sentence	BLEU	SBERT
Original English Translation	故殿おはせましかば、ここなる人々も、 かかるすさびごとにぞ、心は乱らまし」とうち泣きたまふ。 (The subject is omitted here.) cried saying " If the late Lord were still alive, I'm sure even the people staying here would have had to worry their heads about these whimsical play things."		
Reference English Translation	故殿が生きていらっしゃったら、こちらの息子たち だってこうした気まぐれな遊び事で頭を 悩ませていたことでしょうに」とお泣きになる。 (The subject is omitted here.) cried saying " If the late Lord were still alive, I'm sure even my sons here would have had to worry their heads about these whimsical play things."		
The parallel data English Translation	亡き殿がご存命でいらっしゃったら、ここにいる 人々も、こうした遊び事に心を 奪われていたでしょうに」とお泣きになる。 (The subject is omitted here.) cried saying " If the late Lord were still alive, I'm sure even the people staying here would have been fascinated by these play things."	35.16	0.83
The title-added parallel data English Translation	亡き殿がご存命でいらっしゃったら、ここにいる 人たちも、こうした慰みごとのために心が乱れ たことでしょう」とお泣きになる。 (The subject is omitted here.) cried saying " If the late Lord were still alive, I'm sure even the people staying here would have been disturbed by these comforts."	28.26	0.74

Table 9: Translation examples using the parallel and title-added parallel data

on Natural Language Processing for Digital Humanities, pages 53–58, Taipei, Taiwan. Association for Computational Linguistics.

Sho Hoshino, Yusuke Miyao, Shunsuke Ohashi, Akiko Aizawa, and Hikaru Yokono. 2014. Machine translation from historical Japanese to contemporary Japanese using parallel corpus. In *Proceedings of the NLP2014, (In Japanese)*, pages 816–819.

Kanako Komiya, Nagi Oki, and Masayuki Asahara. 2022. Word sense disambiguation of corpus of historical Japanese using Japanese BERT trained with contemporary texts. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 438–446, Manila, Philippines. De La Salle University.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Andrew Piper and Matt Erlin. 2022. The predictability of literary translation. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 155–160, Taipei, Taiwan. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

SHOGAKUKAN. 2010. *The Complete Collection of Japanese Classical Literature (New Edition)*. SHOGAKUKAN.

Masashi Takaku, Tosho Hirasawa, Mamoru Komachi, and Kanako Komiya. 2020. Neural machine translation from historical Japanese to contemporary Japanese using diachronically domain-adapted word embeddings. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 534–541, Hanoi, Vietnam. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need.

Francis Zheng, Edison Marrese-Taylor, and Yutaka Matsuo. 2022. A parallel corpus and dictionary for Amis-Mandarin translation. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 79–84, Taipei, Taiwan. Association for Computational Linguistics.

Measuring the distribution of Hume's Scotticisms in the ECCO collection

Iiro Tiihonen **Aatu Liimatta** **Lidia Pivovarova** **Tanja Säily** **Mikko Tolonen**
University of Helsinki
first.last@helsinki.fi

Abstract

This short paper studies the distribution of Scotticisms from a list compiled by David Hume in a large collection of 18th century publications. We use regular expression search to find the items on the list in the ECCO collection, and then apply regression analysis to test whether the distribution of Scotticisms in works first published in Scotland is significantly different from the distribution of Scotticisms in works first published in England. We further refine our analysis to trace the influence of variables such as publication date, genre and author's country of origin.

1 Introduction

The 18th century was a period of standardization efforts for the English language, which is reflected in many contemporary texts, including those of philosopher David Hume. One of the widely discussed topics was "Scotticisms", i.e. non-standard words and expressions of Scottish origin (Dossena, 2005, 2012).

Hume, himself a Scot, was concerned with the purity of language and paid a great deal of attention to the language of his writings. He was also involved in editorial work and assisted other writers in polishing their texts. The matter of Scotticisms is mentioned several times in his letters to other writers while discussing their work.

A list of "Scotticisms" was published as an appendix to Hume's *Political Discources* in 1752. The list was reprinted several times during the 18th century and mentioned in works by various authors. However, neither the 1752 edition nor any other work by Hume explains how the list was compiled, to what extent it is representative of the language use in 18th century and what the impact of this work was on further standardization efforts. These characteristics make it an interesting cultural artifact to study quantitatively. It might capture regional, generational, genre-related and other variation and eventual standardization of English, but it is also most likely affected by the experiences and interests of its famous author. For example, previous research has found out that many of Hume's Scotticisms occurred in legal contexts (Cruickshank, 2013), which might be explained by the fact that Hume as a former student of law had been exposed to them. Better understanding of how the use of Hume's Scotticisms varied in eighteenth-century texts can provide insights both to the standardization and variation of historical English, and to the origins, nature and limitations of the list itself.

We perform a large-scale corpus study to determine: (i) How Scottish were Hume's "Scotticisms"? (ii) Who used them and where? (iii) Was there change over time, did efforts like Hume's make a difference? To that end, we search for the Scotticisms on the list in a corpus of 18th century publications from England and Scotland and study their distribution across location, genre and time.

There was a previous attempt to search the items on Hume's list in a limited correspondence corpus (Cruickshank, 2013). However, as far as we are aware, this is the first attempt to analyse the actual usage of these items in a massive database of public discourse.

2 Data

The main textual dataset for our analysis is *Eighteenth Century Collections Online* (ECCO) (Tolonen et al., 2021). The texts in ECCO have been made into a machine-readable form using optical character recognition (OCR) technology. However, the ECCO dataset has significant problems with the quality of the OCR texts. Since the dataset itself consists of bitonal microfilm scans of varying quality, the OCR process has often not been able to reproduce the text very well or at all. Thus,

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 36–44
December 1-3, 2023 ©2023 Association for Computational Linguistics

the textual data is often very messy, which can cause problems with analyses of the data (Hill and Hengchen, 2019). Nevertheless, a lot of the ECCO data is of fair quality, and as such the dataset can be used for various kinds of analyses.[1]

Comprising about 200,000 volumes of 18th-century printed works, the ECCO dataset covers roughly half of the surviving printed works from the period. The dataset is not a balanced linguistic corpus, but more an incidental collection of various texts. For instance, ECCO contains many more documents from the later periods of the century, with earlier periods underrepresented in relative terms. Furthermore, different document lengths can dominate different periods. The dataset has not been balanced with respect to other variables, such as genre or register. Moreover, ECCO includes multiple editions of many works, which can confuse quantitative analyses due to the same or very similar content being included multiple times at different time periods.

Our metadata for the texts originates from the *English Short Title Catalogue* (ESTC)[2], harmonized and augmented by the Helsinki Computational History Group (Lahti et al., 2019). Metadata were used to implement two data-filtering steps intended to control for the complexity of the analysis and data quality issues:

- pamphlets and editions other than first were excluded;
- editions must have been published between 1700 and 1799;

And additional steps for the regression analysis:

- the median OCR quality of the pages in the edition had to be at least 80%;
- only editions by authors who were born between 1630–1780 and had at least 5 editions in the ESTC were used;
- the word "Scotticism(s)" must not be mentioned in the text of the edition, as it indicates a linguistic work discussing Scotticisms.

3 Method

3.1 Scotticism Extraction

We retrieved Hume's list of Scotticisms in plaintext format from the Lexicons of Early Modern English (LEME).[3] To identify Hume's Scotticisms in the texts in ECCO, we operationalized the Scotticisms on Hume's list as regular expressions, which enables us to look for various alternative versions of the Scotticism in question. For example, in the case of *cause him do it*—which Hume says should be *cause him to do it* instead—there are two major varying components. First, we included different conjugations of the verb *to cause*, such as *causes*, *caused*, or *causing*. Second, we included other object pronouns, such as *her*, *me*, and *us*, in addition to the pronoun *him*. Furthermore, we accounted for varying spelling conventions by including some common variant spellings, such as *caus'd* and *causd* for *caused*. Finally, cognizant of the varying OCR quality of the data, we included provisions for some common OCR errors, most importantly for the tendency of the long *s* being erroneously recognized as an *f*. After these considerations, and judging that the final *it* is not central to the structure, the regular expression lookup for *cause him do it* was the following:

```
cau(s|f)(es|ed|e|'d|d|ing) (me|you|him|her|it|us|them) do
```

For other items, where necessary, we also considered other potential varying components such as plural and singular forms for nouns, different determiners, and multi-word expressions which could be written separately, together, or hyphenated.

While these kinds of lookups using regular expressions do increase the recall, they do not find all possible instances of the relevant construction. For example, in the case of *cause him do it*, the object *him* could also be any noun phrase, and *do* could be any verb. However, identification of such constructions would require part-of-speech tagging and structural parsing of the ECCO corpus. This procedure by itself may introduce additional errors and would require estimation of tagging and parsing performance of existing tools for historical OCRed data, which is out of the scope of this paper. Similarly, some items on Hume's list of Scotticisms would only be possible to identify using parsed data, and therefore had to be excluded from our analysis.

For some items, Hume speaks against the use of

[1] ECCO is distributed by the Gale company: https://www.gale.com/intl/primary-sources/eighteenth-century-collections-online. Many research groups working with 18th century British data have obtained the license since this is the primary data source for this period.

[2] http://estc.bl.uk/F/?func=file&file_name=login-bl-estc

[3] https://leme.library.utoronto.ca/

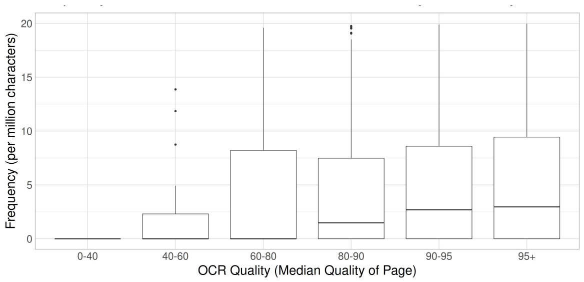


Figure 1: The frequency of Scotticisms in relation to OCR quality of the documents (specified by the collection distributor, i.e. Gale). 364 observations omitted from the visualization due to failing outside visualization range.

a more widely used word in a specific sense. For instance, while *chimney* is widely used in English to refer to a smokestack, Hume rejects its Scottish use to refer to the grate under a fire in a fireplace. As semantic disambiguation would be required to identify the word being in the Scottish sense while ignoring its use in other senses, such items were excluded from our analysis. In total, we were able to identify at least to some degree 67 Scotticisms out of the 106 items on Hume's list. The list of extracted Scotticisms and corresponding regular expressions is presented in Appendix B.

Since the OCR process often leads to changes in spelling and to other forms of textual 'noise', it is possible that we fail to find many occurrences of Scotticisms. Our analysis of the relation between OCR quality and observed Scotticisms in a document is illustrated in Figure 1. Low level of OCR quality is associated with a lower frequency of Scotticisms. Hence we hypothesize that the real differences in the use of Scotticisms in 18th-century Britain might have been even higher than what we have measured. The figure tentatively suggests that median OCR page quality of below 60 percent leads to a drastic drop in observed Scotticisms, and from there on the number of Scotticisms in a typical edition increases from 0 for the 60–80 percent OCR quality bracket to around 3 per million characters for the 90–95 percent bracket.

3.2 Regression Analysis

We analysed the relation of Hume's Scotticisms to other characteristics of publications with *multivariate regression analysis*. This was done to verify which factors related to Scotticisms would hold when other potentially related variables were also controlled for.

Univariate analyses would suggest that a higher average rate of Scotticisms in a group of works tends to correlate with a much higher variation in their number, a phenomenon not captured by many standard regression models. We also suspected that there might be more zero-Scotticism instances than standard statistical models assume. And, by making use of the publisher and author information, our data became relatively high-dimensional, creating the risk of overfitting. We need to assess uncertainty in the model while making it sufficiently complex to incorporate all of these properties. Our solution was to implement a Bayesian zero-inflated Negative Binomial regression model with the R-package BRMS (Bürkner, 2017).[4] Four chains of 3,000 iterations (half of these samples were warm-ups for each chain) were run on STAN (Stan Development Team, 2023) via BRMS to obtain an approximation of the posterior distribution.

Negative Binomial distribution allows us to model the significant increase of variance as a function of the mean (heteroscedasticity), zero-inflation addresses the problem of possible over-representation of zeroes, and by setting a horseshoe-prior to the fixed effects, we can guide the model to by default, e.g. in absence of significant evidence in the data, be in favour of considering none of the effects to matter, which most likely is the case. Similarly, it is easy to group variables together in the Bayesian framework, making them share information in the model fitting process. And, as the posterior distribution is simulated in our Bayesian approach, the model fitting process also produces estimates about the reliability of our findings (effect sizes).

In total, 8,948 editions were used for the regression analysis. Of these, 8,000 were used to fit the model, and the rest were reserved for model evaluation. 812 of the observations reserved for evaluation had an author that was also present in the data used to fit the model, making predictions—and hence the kind of evaluations conducted—with the Bayesian model possible.

We modeled the relation of the number of Scotticisms in an edition to its metadata features. The length of the book in characters was used as an offset to normalize for the length of the book. We included two types of variables in the model. Population-level variables affected all observations with an equal impact. Hence, population-level vari-

[4] `https://CRAN.R-project.org/package=brms`

ables can be associated with fixed effects of regular linear models, as their effect (e.g. that of the OCR quality) only varies by the value of the variable itself, not by some other variable. The impact of the group-level effects varied by a grouping variable. That is to say that they (e.g. the time of publication of an edition) affected the target variable (Scotticisms) differently based on the value of some other variable(s) (e.g. place and genre of the publication) called the grouping variable(s). In our case, the group-level effects are a constant grouped by the author and a quarter-century specific effect grouped by the publication place and genre. In other words, we model the Scotticisms as being affected by a constant unique to each author and progress of time that was conditioned by the combination of genre and time: e.g., Scottish law having its own temporal trajectory.

4 Results

We evaluated the reliability of our model by comparing how well it predicted the frequency of Scotticisms in a test set compared to a null model that only considers the document length in characters to predict a number of Scotticisms.[5] The comparison of the predictions of the model to the real number of Scotticisms in the test data is shown in Figure 2. There is a clear difference in favour of the full model, as it is able to more accurately detect those instances in which the number of Scotticisms in a document is very high. Our model is able to capture such variation in the number of Scotticisms that generalizes beyond the training set.

The results of the regression analysis are presented in Table 1, as well as in Tables 2 and 3 in Appendix A. These tables report the approximated marginal posterior-distributions of the model. They communicate estimates of how well the whole range of possible parametrizations of the model explain the data and align with priors using posterior-likelihood as the measure. The probabilities related to any given parameter values express how big a proportion of the posterior-likelihood (compared to all possible parametrizations) is concentrated on those parameter values. For example, if some parameter's (e.g. the effect of OCR quality) marginal posterior distribution at 2.5th quantile is 0.1 and at 97.5th quantile 0.2, we can say that 95 percent of the posterior-likelihood (or posterior-probability) is concentrated on models that propose that the parameter is positive and between 0.1 and 0.2. The tables allow us to identify those parameters for which most (95 percent) of the posterior-probability is supporting either a positive or a negative effect on Scotticisms.

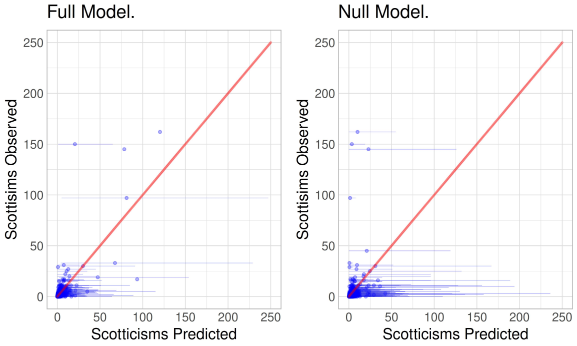


Figure 2: The real and predicted (mean) number of Scotticisms as predicted by the full and null models. When possible within the limits of the x-axis, the range of predictions from 2.5th to 97.5th quantile is illustrated with blue horisontal lines. 3 observations omitted from the full and 1 from the null model due to zooming of the image.

Several findings of the preceding univariate visualizations are supported by the posterior distribution of the fitted model:[6] good OCR quality, the author being a Scot born in the 17th century and especially Scottish legal publications are associated with an increased rate of Scotticisms. Additionally, several authors are associated with an increased rate of Scotticisms. The most consistent factors associated with a lower rate of Scotticisms are those depicting the difference between the first and 3rd/4th quarters of the 18th century for non-legal Scottish publications.

Hence, the regression model offers support for three major claims:

1. For genres other than law, the rate of Hume's Scotticisms decreased in Scottish publications during the later 18th century.
2. Scottish legal publications used Hume's Scotticisms at a much higher rate than other types of publications.
3. Author-to-author variation in the use of Scotticisms was significant, and the 17th-century Scottish generations used them more.

Based on these claims, we can draw two higher-

[5] The number of iterations and chains for the null model was the same as for the full model.

[6] Here, we only discuss such effects for which the tails (2.5th, 97.5th quantiles) of the posterior are both either positive or negative. That is, we focus on effects that the model sees as very likely positive or very likely negative.

Variable	2.5th q.	Median	97.5th q.
Intercept	0.68	1.9	3.7
OCR quality	0.094	0.13	0.16
Scot a. 17th c.	0.007	0.53	0.85
Other a. 18th c.	-0.6	-0.24	-0.0001

Table 1: Posterior distributions of population-level effects. Includes only those effects for which the sign of the 2.5th and 97th quantiles was the same (i.e., the effect is highly likely either positive or negative).

level conclusions. First, the overall process of standardization was best resisted by the often formulaic legal genre of Scottish texts, which did not show robust signs of decreased use of Scotticisms even at the end of the century. In previous research it has been suggested that Scottish legal language got replicated throughout society because law had a daily impact on the lives of most Scots (Cruickshank, 2013, 39). Our results imply the opposite: while legal texts remained Scotticism-heavy, literary culture as a whole was heavily impacted by standardization.

The other major conclusion is that individuals differed in their use of Scotticisms to a remarkable degree. Even after accounting for other factors that were related to variation in the use of Scotticisms (among them the overall difference between authors born in the seventeenth vs. the eighteenth century), some authors used them orders of magnitude more than others. While the analysis of the use of Scotticisms by specific individuals is beyond the scope of this paper, this differentiation as a general phenomenon has historical implications. Hume and fellow-minded advocates of standardization were focusing on removing characteristics of English that did divide authors.

It is worth noting that most items from Hume's list were more frequently used in their standard form, even in Scottish writings. For example, the Scotticism *alwise* was found 173 times in our corpus of works published in Scotland, while the standard form *always* was used 44,972 times in this corpus. Thus, *alwise* could serve as a strong predictor of Scottish work, but even in Scottish works the standard form was much more frequent. Therefore, the standardization was not a complete transformation of Scottish English but an attempt to eliminate what was seen as regionally specific language mistakes.

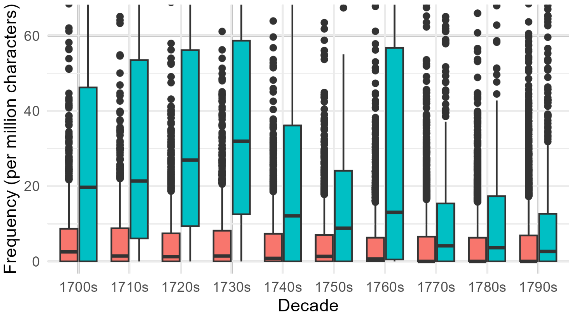


Figure 3: Frequency of Scotticisms per decade in works published in England (red) and Scotland (blue).

Taken as a whole, the changes in the use of Scotticisms were remarkable. Figure 3 shows that Scotticisms are indeed prevalent in books published in Scotland and that even there their number gradually decreases during the 18th century, with the 1760s being an outlier. We checked those documents from 1760s that have the biggest number of Scotticisms and found out that all of them were legal documents. The peak in the number of Scottish legal documents published during that time could have some historical explanation or could be a mere corpus artefact. We leave this for further investigation.

5 Conclusion and Further Work

Our analysis confirms that David Hume was familiar with the peculiarities of language use in his time. The overall trend towards standardization was resisted by Scottish legal texts and modified by significant variation between authors. The specific contribution of David Hume to this process is a matter for further research.[7]

Further work would include both refining the methods and taking into account a broader set of materials. The former line of research may include structural analysis of the data—including part of speech tagging and parsing—and efforts to find "Scotticisms" in a data-driven way without any predefined list. The latter would involve studying Scotticisms compiled by other 18th century writers, and studying other types of data outside ECCO, such as newspapers and personal correspondence. We also plan to apply methods based on contextualized embeddings for semantic disambiguation and post-OCR correction of items from the list.

Acknowledgments

The research was supported by the Academy of Finland under the projects *Rise of commercial society*

[7] A paper discussing these results from a humanities perspective is currently under review (Tolonen et al., 2024).

and eighteenth-century publishing (grant numbers 333716 and 333717) and *High performance computing for the detection and analysis of historical discourses* (grant numbers 347706 and 347709). Computational resources were provided by CSC—IT Center for Science.

References

Paul-Christian Bürkner. 2017. BRMS: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28.

Janet Cruickshank. 2013. The role of communities of practice in the emergence of Scottish Standard English. In Joanna Kopaczyk and Andreas H. Jucker, editors, *Communities of Practice in the History of English*, pages 19–45. John Benjamins, Amsterdam.

Marina Dossena. 2005. *Scotticisms in Grammar and Vocabulary: 'Like Runes Upon a Standin' Stane'?* John Donald, Edinburgh.

Marina Dossena. 2012. Scottishness and the book trade: Print and Scotticisms. In Warren McDougall and Stephen W. Brown, editors, *The Edinburgh History of the Book in Scotland*, volume 2, pages 545–550. Edinburgh University Press, Edinburgh.

Mark Hill and Simon Hengchen. 2019. Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. *Digital Scholarship in the Humanities*, 34(4):825–843.

Leo Lahti, Jani Marjanen, Hege Roivainen, and Mikko Tolonen. 2019. Bibliographic data science and the history of the book (c. 1500–1800). *Cataloging & Classification Quarterly*, 57(1):5–23.

Stan Development Team. 2023. *Stan Modeling Language User's Guide and Reference Manual*. Version 2.31.

Mikko Tolonen, Aatu Liimatta, Lidia Pivovarova, Iiro Tiihonen, and Tanja Säily. 2024. Hume's list of Scotticisms in eighteenth-century British context.

Mikko Tolonen, Eetu Mäkelä, Ali Ijaz, and Leo Lahti. 2021. Corpus linguistics and Eighteenth Century Collections Online (ECCO). *Research in Corpus Linguistics*, 9(1):19–34.

A Regression Analysis Results

Variable	2.5th q.	Median	97.5th q.
Law England (intercept)	-2.04	-0.24	1.05
Law Scotland (intercept)	0.99	2.94	4.80
Other England (intercept)	-2.14	-0.39	0.82
Other Scotland (intercept)	-0.30	1.51	2.76
Religion and Philosophy England (intercept)	-1.93	-0.17	1.04
Religion and Philosophy Scotland (intercept)	-0.41	1.39	2.65
Law England 2nd quarter	-1.61	-0.83	-0.02
Law Scotland 2nd quarter	-1.77	-0.44	0.53
Other England 2nd quarter	-0.24	-0.07	0.09
Other Scotland 2nd quarter	-0.53	-0.02	0.48
Religion and Philosophy England 2nd quarter	-0.30	-0.13	0.02
Religion and Philosophy Scotland 2nd quarter	-0.81	-0.36	0.04
Law England 3rd quarter	-1.03	-0.29	0.38
Law Scotland 3rd quarter	-0.90	0.22	1.20
Other England 3rd quarter	-0.39	-0.20	-0.03
Other Scotland 3rd quarter	-1.78	-1.30	-0.83
Religion and Philosophy England 3rd quarter	-0.52	-0.33	-0.15
Religion and Philosophy Scotland 3rd quarter	-1.30	-0.88	-0.46
Law England 4th quarter	-0.75	-0.13	0.46
Law Scotland 4th quarter	-2.09	-0.85	0.21
Other England 4th quarter	-0.30	-0.12	0.07
Other Scotland 4th quarter	-1.95	-1.44	-0.97
Religion and Philosophy England 4th quarter	-0.44	-0.24	-0.04
Religion and Philosophy Scotland 4th quarter	-1.56	-1.13	-0.71

Table 2: The effect of different combinations of genre and place on the rate of Scotticisms (intercept) and how the effect changes in the 2nd, 3rd, and 4th quarters of the 18th century.

Variable	2.5th q.	Median	97.5th q.
Hamilton, James Hamilton, Duke of, 1724-1758.	2.01	2.70	3.45
Palmer, Thomas Fyshe, 1747-1802.	1.89	2.53	3.25
Law, William, 1686-1761.	1.50	2.08	2.72
Cullen, Francis Grant, Lord, 1658-1726.	1.40	2.05	2.80
Cardonnel, Adam de, -1820.	1.35	2.01	2.74
Simson, John, 1668?-1740.	1.27	1.97	2.76
Mackenzie, Alexander, 1735-1805.	1.30	1.84	2.41
Kirkby, John, 1705-1754.	0.99	1.78	2.63
Aberdeen, George Gordon, Earl of, 1722-1801.	1.08	1.77	2.52
Blackstone, William, Sir, 1723-1780.	0.89	1.66	2.50
Mitford, William, 1744-1827.	0.87	1.56	2.36
Maittaire, Michael, 1667-1747.	0.95	1.56	2.26
Roscoe, William, 1753-1831.	0.90	1.52	2.19
Howie, John, 1735-1793.	0.86	1.51	2.26
Roxburghe, John Ker, Duke of, 1740-1804.	0.71	1.51	2.38
Cockman, Thomas, 1675?-1745.	0.85	1.47	2.17
Badeslade, Thomas.	0.74	1.47	2.30
Okely, Francis, approximately 1719-1794.	0.66	1.47	2.27
Cockburn, William, Sir, 1662-1751.	0.72	1.43	2.20
Baretti, Giuseppe, 1719-1789.	0.77	1.41	2.13
Eachard, John, 1636?-1697.	0.71	1.39	2.12
Moray, James Stuart, Earl of, 1708-1767.	0.57	1.38	2.34
Gib, Adam, 1714-1788.	0.72	1.36	2.01
Guyon, Jeanne Marie Bouvier de La Motte, 1648-1717.	0.73	1.35	2.00
Middleton, Conyers, 1683-1750.	0.88	1.31	1.77
Robe, James, 1688-1753.	0.64	1.31	2.07
Forrester, Thomas, 1635?-1706.	0.54	1.31	2.16
Anderson, James, 1739-1808.	0.71	1.30	1.92
Grove, Henry, 1684-1738.	0.85	1.27	1.72
Coote, Charles, 1761-1835.	0.74	1.21	1.73
Heathcote, Ralph, 1721-1795.	0.51	1.19	1.88
Tucker, Abraham, 1705-1774.	0.60	1.19	1.81
Jackson, John, 1686-1763.	0.71	1.18	1.69
Brown, John, 1722-1787.	0.75	1.14	1.53
Ireland, Samuel, -1800.	0.50	1.13	1.77
Hare, Francis, 1671-1740.	0.58	1.09	1.62

Table 3: Authors with the highest “random” effect on the rate of Scotticisms.

B Scotticisms and Regular Expressions

<table>
<tr><th>Scotticism</th><th>Regular expression</th></tr>
<tr><td>conform to</td><td>conform(s|ed|’d|d|ing)? to</td></tr>
<tr><td>friends and acquaintances</td><td>friends and acquaintances</td></tr>
<tr><td>maltreat</td><td>maltreat(s|ed|’d|d|ing)?</td></tr>
<tr><td>advert to</td><td>advert(s|ed|’d|d|ing)? to</td></tr>
<tr><td>proven</td><td>proven</td></tr>
<tr><td>improven</td><td>improven</td></tr>
<tr><td>approven</td><td>approven</td></tr>
<tr><td>pled</td><td>pled</td></tr>
<tr><td>incarcerate</td><td>incarcerat(es|ed|e|’d|d|ing)</td></tr>
<tr><td>fresh weather</td><td>fre(s|f)h weather</td></tr>
<tr><td>in the long run</td><td>in the long run</td></tr>
<tr><td>notwithstanding of that</td><td>notwith(s|f)tanding of that</td></tr>
<tr><td>a question if</td><td>a que(s|f)tion if</td></tr>
<tr><td>with child to a man</td><td>with child to a man</td></tr>
<tr><td>simply impossible</td><td>(s|f)imply impo(s|f)(s|f)ible</td></tr>
<tr><td>in time coming</td><td>in time coming</td></tr>
<tr><td>nothing else</td><td>nothing el(s|f)e</td></tr>
<tr><td>nothing else</td><td>no thing el(s|f)e</td></tr>
<tr><td>severals</td><td>(s|f)everals</td></tr>
<tr><td>anent</td><td>anent</td></tr>
<tr><td>allenarly</td><td>allenarly</td></tr>
<tr><td>alongst</td><td>along(s|f)t</td></tr>
<tr><td>as I shall answer</td><td>as I (s|f)hall an(s|f)wer</td></tr>
<tr><td>cause him do it</td><td>cau(s|f)(es|ed|e|’d|d|ing)(me|you|him|her|it|us|them) do</td></tr>
<tr><td>marry upon</td><td>marr(ying|y’d|ies|ied|yd|y) upon</td></tr>
<tr><td>effectuate</td><td>effectuat(es|ed|e|’d|d|ing)</td></tr>
<tr><td>a wright</td><td>a wright</td></tr>
<tr><td>defunct</td><td>defunct</td></tr>
<tr><td>evite</td><td>evit(es|ed|e|’d|d|ing)</td></tr>
<tr><td>part with child</td><td>part(s|ed|’d|d|ing)? with child</td></tr>
<tr><td>notour</td><td>notour</td></tr>
<tr><td>to be difficulted</td><td>(am|is|are|was|were|been|being|be) difficult(ed|’d)</td></tr>
<tr><td>think shame</td><td>(thinking|thinks|think|thought) (s|f)hame</td></tr>
<tr><td>in favours of</td><td>in favou?rs of</td></tr>
<tr><td>dubiety</td><td>dubiet(ys|y’s|y|ies)</td></tr>
<tr><td>compete</td><td>compet(es|ed|e|’d|d|ing)</td></tr>
<tr><td>remeed</td><td>reme(de|ed|id|ad)(s|ed|’d|d|ing)?</td></tr>
<tr><td>bankier</td><td>bankier(s’|’s|s)?</td></tr>
<tr><td>adduce a proof</td><td>adduc(es|ed|e|’d|d|ing) a proof</td></tr>
<tr><td>superplus</td><td>(s|f)uper−?plu(s|f)((s|f)?es)?</td></tr>
<tr><td>forfaulture</td><td>forfaulture(s’|’s|s)?</td></tr>
<tr><td>in no event</td><td>in no event</td></tr>
<tr><td>common soldiers</td><td>common (s|f)oldier(s’|’s|s)?</td></tr>
<tr><td>debitor</td><td>debitor(s’|’s|s)?</td></tr>
<tr><td>exeemed</td><td>exee?m(ed|’d|d)</td></tr>
<tr><td>yesternight</td><td>ye(s|f)ternight</td></tr>
<tr><td>big coat</td><td>big coat(s’|’s|s)?| big−?coat(s’|’s|s)?</td></tr>
<tr><td>tenible argument</td><td>tenible argument(s’|’s|s)?</td></tr>
<tr><td>amissing</td><td>a−?mi(s|f)(s|f)ing</td></tr>
<tr><td>extinguish an obligation</td><td>extingui(s|f)h(es|ed|’d|d|ing)? (an|the|(my|your|his|her|its|our|their)) obligations? | extingui(s|f)h(es|ed|’d|d|ing)? obligations?</td></tr>
<tr><td>depone</td><td>depon(es|ed|e|’d|d|ing)</td></tr>
<tr><td>to inquire at a man</td><td>(e|i)nquir(es|ed|e|’d|d|ing) at a (man|person)</td></tr>
<tr><td>angry at</td><td>angry at</td></tr>
<tr><td>to send an errand</td><td>(s|f)en(ding|ded|ds|d|t) (an|the) | (s|f)en(ding|ded|ds|d|t) errands?</td></tr>
<tr><td>to furnish goods to him</td><td>furni(s|f)h(es|ed|’d|d|ing)? goods to (me|you|him|her|it|us|them)</td></tr>
<tr><td>to open up</td><td>open(s|ed|’d|d|ing)? up</td></tr>
<tr><td>Thucydide</td><td>thucydide</td></tr>
<tr><td>Herodot</td><td>herodote?</td></tr>
<tr><td>Sueton</td><td>sueton</td></tr>
<tr><td>butter and bread</td><td>butter and bread</td></tr>
<tr><td>pepper and vinegar</td><td>pepper and vinegar</td></tr>
<tr><td>paper, pen and ink</td><td>paper,? pen,? and ink</td></tr>
<tr><td>as ever I saw</td><td>as ever (I|you|he|she|it|we|they) saw</td></tr>
<tr><td>come in to the fire</td><td>(comes|come|coming|came) in to the fire</td></tr>
<tr><td>alwise</td><td>alwi(s|f)e</td></tr>
<tr><td>cut out his hair</td><td>(cut|cuts|cutting) out (my|your|his|her|its|our|their) hair</td></tr>
<tr><td>to get a stomach</td><td>(gotten|getting|get|got) a stomach (for|to)</td></tr>
<tr><td>vacance</td><td>vacance(s’|’s|s)?</td></tr>
</table>

Table 4: Scotticisms from Hume’s list and regular expressions used to find them in ECCO.

Effect of data quality on the automated identification of register features in Eighteenth Century Collections Online

Aatu Liimatta
University of Helsinki
aatu.liimatta@helsinki.fi

Abstract

Many large-scale investigations of textual data are based on the automated identification of various linguistic features. However, if the textual data is of lower quality, automated identification of linguistic features, particularly more complex ones, can be severely hampered.

Data quality problems are particularly prominent with large datasets of historical text which have been made machine-readable using optical character recognition (OCR) technology, but it is unclear how much the identification of individual linguistic features is affected by the dirty OCR, and how features of varying complexity are influenced differently.

In this paper, I analyze the effect of OCR quality on the automated identification of the set of linguistic features commonly used for multidimensional register analysis (MDA) by comparing their observed frequencies in the OCR-processed *Eighteenth Century Collections Online* (ECCO) and a clean baseline (ECCO-TCP). The results show that the identification of most features is disturbed more as the OCR quality decreases, but different features start degrading at different OCR quality levels and do so at different rates.

1 Introduction

Large-scale textual datasets have become increasingly common in computational linguistics and in various subfields of digital humanities. Naturally, such datasets cannot easily compare to the high quality of traditional (but much smaller) linguistic corpora, which have been carefully curated for balance and manually edited to ensure the accuracy of the textual data to as large a degree as feasible. Instead, it is the expectation that when analyzed in aggregate, the large amount of data can smooth over many of the flaws which would prove very difficult to work around with smaller datasets or when focusing the analysis on individual texts or individual instances of items. Even then, the overall lower quality of the data causes issues for many linguistic and other digital humanities analyses, such as in the case of, for example, the analysis of social media data (e.g. Eisenstein, 2013).

However, a major source of large-scale low-quality textual humanities data, particularly in fields such as historical linguistics and computational history, are texts which have been turned from scans of physical documents into machine-readable format using optical character recognition (OCR) technology. For instance, the OCR quality of *Eighteenth Century Collections Online* (ECCO)[1], a collection of over 200,000 mostly English-language works published in the United Kingdom during the 18th century (see Tolonen et al., 2021) and a central resource for DH scholars (Gregg, 2020), is extremely variable. A main contributing factor of the often low OCR quality for ECCO is the OCR being run on bitonal scans of microfilms with OCR algorithms which have not been fine-tuned for eighteenth-century typefaces or trained to recognize e.g. the long *s* character ⟨ſ⟩.

Earlier studies on the effects of OCR errors in ECCO (e.g. Hill and Hengchen, 2019) are largely focused on individual tokens, characters, and n-grams. In contrast, the present study focuses on the effect of the OCR errors in ECCO on the automated identification of a set of more complex linguistic features commonly used for the multi-dimensional method of register analysis (MDA) (see e.g. Biber, 1988; Biber and Conrad, 2009).

[1] https://www.gale.com/intl/primary-sources/eighteenth-century-collections-online

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 45–51
December 1-3, 2023 ©2023 Association for Computational Linguistics

2 Background

2.1 OCR and ECCO

The difficulties caused by dirty OCR have been long recognized in the literature (e.g. Lopresti, 2009; Traub et al., 2015; Vitman et al., 2022). When it comes to the ECCO dataset, Hill and Hengchen (2019) compare the ECCO-TCP[2], a manually keyed version of a subset of the documents included in ECCO, to the same set of documents from the regular OCR-processed ECCO (henceforth: ECCO-OCR) both on the basis of token and type similarity and in a number of bag-of-words approaches used in digital humanities, such as topic modeling and methods of authorship attribution. They find that, for example, the mean OCR precision in their dataset is 0.744, meaning that on the average page, 74% of the tokens are correct, whereas the recall is 0.814, meaning that 81% of the tokens are included in the OCR version.

2.2 Multi-dimensional analysis

MDA, originally developed by Biber (1988), is a corpus-linguistic approach which extracts functional dimensions from a textual dataset; the dimensions describe variation in the communicative purposes and situational concerns between the texts within the dataset (see e.g. Biber, 1988; Biber and Conrad, 2009), each dimension comprising a gradient between two poles with opposing functions.

The central idea behind the MDA methodology is that linguistic features which are better-suited to the function and situational concerns of a text are more likely to be used in the text. Consequently, commonly co-occurring linguistic features can be assumed to share an underlying set of functions. MDA uses statistical methods such as factor analysis on a set of texts to extract co-occurring (and complementary) groups of features which are then interpreted in terms of their function, forming dimensions of register variation.

To give a basic example, *past tense* verb forms and *third-person pronouns* are naturally more common in narrative contexts than in non-narrative contexts. One of the central findings of Biber (1988) is the gradient between "involved" and "informational" production, with the informational pole characterized by features such as *nouns*, *prepositions*, and *attributive adjectives*, and the complementary involved pole by features such as *private verbs*[3], *THAT deletion*, and *contractions*.[4]

While in principle different sets of linguistic features can be and have been used for MDA analyses, it is common to build a MDA feature set on the core set of linguistic features originally compiled by Biber (1988) through an extensive survey of previous linguistic literature.

2.3 Multi-dimensional analysis and ECCO

Many of the features included in the MDA core set of features are much more specific and more complex than those analyzed by Hill and Hengchen (2019). It is a statistically reasonable assumption that a random OCR error is more likely to occur in a longer multi-word construction than in an individual token. Consequently, it could be expected that dirty OCR would make it more difficult to identify such complex features in the data, and that therefore any analysis making use of such features would be severely disturbed or completely prevented if the set of texts being analyzed contains many OCR errors.

In order to test this assumption, Liimatta et al. (2023) evaluate the effects of dirty OCR on the MDA methodology by comparing the results of the analysis run on ECCO-TCP and separately on the parallel set of documents from ECCO-OCR. Perhaps surprisingly, the MDA dimensions Liimatta et al. (2023) acquire from ECCO-TCP and ECCO-OCR turn out to be very similar, which suggests that even if not every instance of each linguistic feature used in the analysis is identified properly in the ECCO-OCR data, enough instances of most features can still be identified for meaningful co-occurrence patterns to be preserved to a degree.

However, it still remains the case that dirty OCR renders many instances of any features of interest unrecognizable by automated processing methods, and as such, there are linguistic features which are better or worse suited for MDA analysis of dirty OCR datasets such as ECCO-OCR and for other analyses using similar approaches. The aim of the present paper is to explore the core set of features used for MDA and see how each of these features is individually affected by the OCR process, to shed light on which kinds of linguistic features can most robustly be used for MDA and other similar analyses, but also more generally to understand

[2] https://textcreationpartnership.org/tcp-texts/ecco-tcp-eighteenth-century-collections-online/

[3] e.g. *think*, *assume*, or *feel*

[4] For the full list of features included in Biber (1988) and their descriptions, see Biber (1988, Appendix II).

the influence of decreasing OCR quality on the integrity of various linguistic structures.

3 Materials and methods

3.1 Data

All of the textual data used in the present analysis is based on ECCO. The *ECCO Text Creation Partnership* (ECCO-TCP) dataset constitutes the clean baseline for the analysis. ECCO-TCP is a manually keyed version of a small subset of the full ECCO dataset. Thanks to the careful editing process, ECCO-TCP, while not perfect, is close in quality to other hand-edited historical datasets in terms of its transcription accuracy, and as such can be considered a clean standard for the included texts (cf. Hill and Hengchen, 2019).

Additionally, in order to be able to estimate the degradation in feature identification caused by dirty OCR when compared to the clean versions of the texts, I have created as a second dataset a parallel subset of the regular OCR-processed ECCO dataset (ECCO-OCR) whose texts match the texts included in ECCO-TCP. Finally, both datasets were tagged for part of speech[5] using spaCy[6].

The OCR quality estimate is based on the confidence levels reported by the OCR engine which Gale, the publisher of ECCO, used to process the texts. The original OCR confidence was calculated on a per-page basis; for the present analysis, the mean of the OCR quality estimate for all of the pages of a work was used as the overall OCR quality of the work.

3.2 Methods

Both datasets were processed with the same feature identification pipeline, which identified the linguistic features of interest using automated means and counted their occurrences in each of the texts in both datasets. The identification of the features was performed using algorithms mainly based on those provided by Biber (1988)[7]. For instance, the algorithm for the feature *WH-clauses* is given by Biber (1988) as follows:

```
PUB/PRV/SUA + WHP/WHO + xxx
(where xxx is NOT = AUX)
```

This means that *WH-clauses* are identified as starting with a word belonging to the classes of public, private, or suasive verbs (e.g. *say*, *believe*, *agree*), followed by a WH pronoun (i.e. *who*, *whose*, *whom*, *which*) or other WH word (e.g. *what*, *when*, *whether*, *why*, *wherever*), followed by a word which is *not* an auxiliary, defined by Biber (1988) to be either a modal verb, or any of the the verbs *to do*, *to have*, or *to be*.[8]

However, the present study uses a different part-of-speech tagset and tokenization scheme than the original study by Biber (1988), which sometimes requires changes or allows for improvements to the algorithms, and consequently the algorithms for each individual feature were not always followed exactly. Furthermore, a handful of the features were excluded from the analysis because their algorithms require manual checking of the results.

After normalizing the observed feature counts to the observed character count[9] of the text, I calculated the proportion of the frequency observed in the ECCO-TCP version of each of the texts which was observed in the ECCO-OCR version of the same text. In other words, because OCR errors will in many cases prevent the implemented algorithms from correctly identifying the features, e.g. because the word form has character errors or because the part of speech has been misidentified, I calculated by how much the observed frequency of each of the features changed from the clean version of the text to the OCR version of the text. This

[5] Automated part-of-speech (POS) tagging is not perfect, and may itself present some problems for the downstream task of automated identification of linguistic features. Furthermore, OCR errors will lower POS tagging quality for the ECCO-OCR dataset. However, POS tagging is a necessary step for the identification of the linguistic features analyzed in the present study, and as such MDA studies even on perfectly clean datasets typically need to make use of imperfect POS tagging. As the aim of the present study is to gauge the effect of OCR quality on the identification of linguistic features, it is reasonable to use a realistic POS tagging as a baseline, and to include the POS tagging quality degradation in the OCR dataset as part of the overall degradation of feature identification caused by dirty OCR.

[6] https://spacy.io

[7] While the MDA methodology has seen extensive use over the years, Biber (1988) still provides the most comprehensive description of the exact patterns by which the individual features can be identified.

[8] For the full list of the features included in the present study, see Appendix A. For the full list of original algorithms and descriptions of the features, see Biber (1988, Appendix II).

[9] While token count or word count would be more typical bases for normalization, I have chosen to use the character count of the text as the normalization basis for problematic OCR data. The dirty OCR often includes many erroneous spaces between strings of characters, which tends to inflate the number of tokens as created by a typical tokenizer. Consequently, preliminary analyses show that the character count of the OCR text, while still wrong, is likely to be proportionally closer to the true character count of the text than the token count is to its true token count.

change was calculated for each text as

$$\frac{f_{ocr}}{f_{tcp}} - 1$$

where f is the normalized frequency of the feature as observed in either the OCR or TCP version of the text; the offset -1 is to have a value of 0 represent no change from TCP to OCR.

4 Results

Figure 1 shows the results of the analysis for each of the features as a function of the OCR quality of the text. Every facet represents a different linguistic feature, with the OCR quality estimate for the text on the vertical axis, higher OCR quality at the top and OCR quality decreasing towards the bottom. The horizontal axis represents the proportional change in the observed frequency of the feature from ECCO-TCP to ECCO-OCR. Because there is a lot of variation in the data, a smoothing trend line[10] was also included for each feature.

In other words, points at 0.0 on the horizontal axis in Figure 1, i.e. on the dark vertical line, represent texts with the exact same frequency in both ECCO-TCP and ECCO-OCR. Such texts are the ideal case, since they have no change in the observed frequency from the clean version to the OCR version of the text, suggesting that all instances of the feature have been correctly identified. However, in practice, as expected, most texts deviate from the ideal. Points to the left of the middle line represent texts which have a lower frequency of the feature in ECCO-OCR than in ECCO-TCP. For instance, in a text at the -0.5 mark, the OCR version has only half the observed frequency of the feature as compared to the clean version. Similarly, the points to the right of the middle line indicate texts with a higher frequency of the feature in ECCO-OCR than ECCO-TCP, with the 0.5 mark meaning a 50% increase in the observed frequency.

Based on Figure 1, it is clear that as the OCR quality decreases, the observed frequency of identification typically also decreases, as evidenced by the trend line tending down and to the left. But different features are also clearly affected differently by the decrease in OCR quality. For some features, the fall towards the left is very direct and rapid, beginning very close to the highest OCR quality texts, meaning that their identification is instantly

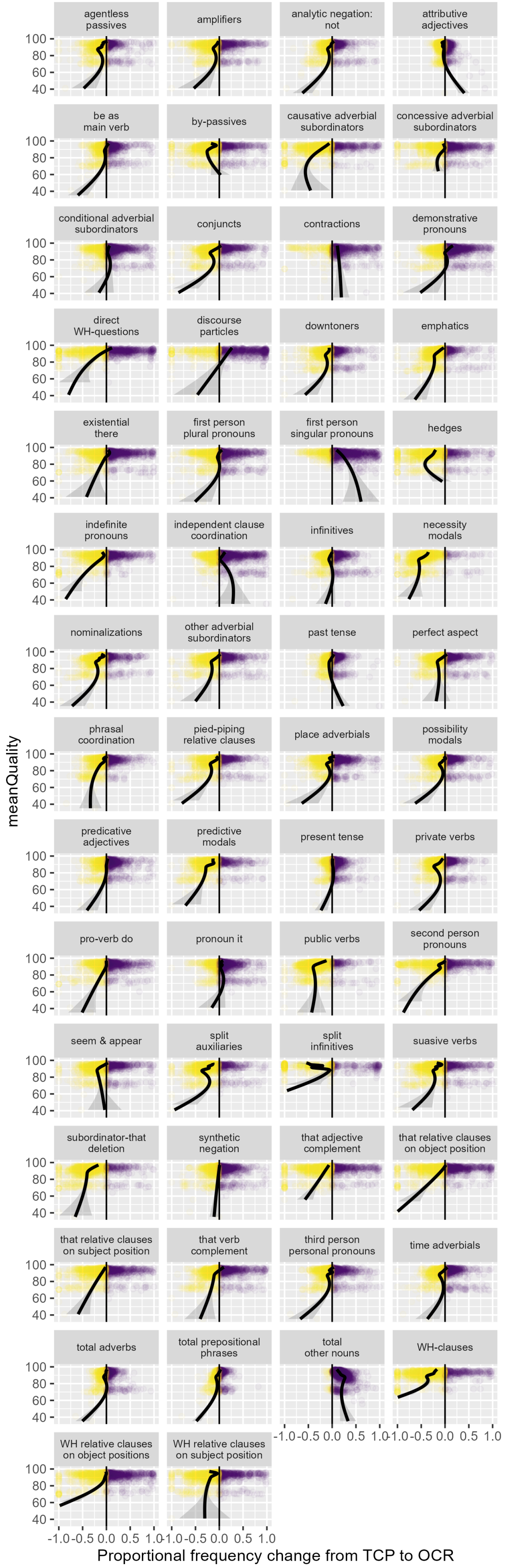


Figure 1: Proportion of ECCO-OCR rate of occurrence of the ECCO-TCP rate of occurrence by OCR quality

[10] ggplot2 (Wickham, 2016): `geom_smooth(method = "gam", formula = y ~ s(x, bs = "cs"))`

affected when the OCR quality decreases. These include features such as *direct WH-questions*[11] and *that relative clauses on object position*[12].

Features which appear to be particularly adversely affected by dirty OCR include *WH-clauses*[13], which have on average lower observed frequencies even in the OCR texts with the best OCR quality, and the frequency drops very rapidly with the OCR quality. Possibly because of OCR problems with the long *s*, *necessity modals*[14] also start at a lower observed frequency even at highest OCR quality, and keep decreasing in frequency with decreasing OCR quality, if somewhat less quickly than WH-clauses do.

In contrast to features of the above type, which start decreasing as soon as the OCR quality decreases, there are a number of features which stay relatively stable for a larger portion of the OCR quality range and only show larger changes in their average observed frequency when the OCR quality drops too much, possibly because they tend to be relatively simple to identify. These features include *present tense* verb forms, which only show a drop below 60-70% OCR quality, *attributive adjectives*[15], which show an increase below a similar OCR quality level, and *predicative adjectives*[16], which show a drop below 70-80% OCR quality. There is also an even larger group of features which do show a small difference from the clean version at higher OCR quality levels but for which this difference is only relatively minor, including e.g. the *analytic negation not*, *first person plural pronouns*[17] and *total adverbs*[18], all of which only show a larger change in their observed frequency below about 70% OCR quality.

Another, rather curious group of features are those whose identified frequencies increase instead of decrease in the OCR dataset. Typically, this increase can be attributed to the OCR process and other processing of the dataset, particularly the tokenization and the part-of-speech tagging. For instance, the *total other nouns*[19] category shows higher frequencies in ECCO-OCR compared to the clean baseline throughout the OCR quality range. This is because the imperfect OCR process creates many strings of characters which the part-of-speech tagger cannot recognize, and most taggers have a tendency of tagging such tokens as nouns, particularly if there are too many unrecognizable tokens in succession to infer a different part of speech from the surrounding structure.

Similarly, *first person singular pronouns* appear to be identified at a close to correct rate in the OCR texts close to the top OCR quality, but there is an increasing number of misidentifications as the OCR quality decreases. This is largely because as the OCR quality degrades, the OCR process produces more and more random *I* characters, which then get misidentified as the first person pronoun *I*.

Finally, there are a few features which appear to be barely affected by the decrease in OCR quality. These are likely relatively simple to identify features which also happen to consist of characters which tend to have been recognized more reliably than average in the OCR process. These features include most prominently *synthetic negation*[20], but features such as *pronoun it* and *infinitives* could also be considered to not be affected very much at all by changes in OCR quality.

5 Conclusion

The results show that, as expected, lower OCR quality leads to lower reliability of identification for most of the features analyzed. However, the effect of OCR on all of the features is not the same, with features which are simpler to identify and less likely to invite OCR errors appearing generally more resistant to lower OCR quality, while features whose identification involves multiple consecutive items from specific lists appear generally more at risk. It is not possible based on this study alone to recommend any single ECCO OCR quality level which could be considered “good enough” for most analyses, as what is good enough depends on the features one is interested in. On the other hand, MDA studies consistently produce “compatible” results regardless of the exact set of features analyzed (McEnery and Hardie, 2012), suggesting that for MDA, focusing on a smaller set of more resilient features may be enough to get better results even from lower-quality textual data.

[11] e.g. *What is that?*

[12] e.g. *the place that they mentioned*

[13] e.g. *I didn't know what it was*

[14] *must*, *should*, and *ought*

[15] e.g. *a blue house*

[16] e.g. *the house is blue*

[17] e.g. *we*

[18] e.g. *quickly*

[19] Nouns not counted as nominalizations.

[20] e.g. *no man*

Acknowledgements

The research was supported by the Academy of Finland under the project *Rise of commercial society and eighteenth-century publishing* (grant number 333717).

References

Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press, Cambridge.

Douglas Biber and Susan Conrad. 2009. *Register, genre, and style*. Cambridge University Press, Cambridge.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369.

Stephen H. Gregg. 2020. *Old Books and Digital Publishing: Eighteenth-Century Collections Online*. Cambridge University Press.

Mark J. Hill and Simon Hengchen. 2019. Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. *Digital Scholarship in the Humanities*, 34(4):825–843.

Aatu Liimatta, Yann Ryan, Tanja Säily, and Mikko Tolonen. 2023. Results from rough data? The large-scale study of early modern historiography with multi-dimensional register analysis. *Digital Humanities in the Nordic and Baltic Countries Publications*, 5(1):297–312.

Daniel Lopresti. 2009. Optical character recognition errors and their effects on natural language processing. *International Journal on Document Analysis and Recognition (IJDAR)*, 12(3):141–151.

Tony McEnery and Andrew Hardie. 2012. *Corpus linguistics: Method, theory and practice*. Cambridge University Press, Cambridge, New York.

Mikko Tolonen, Eetu Mäkelä, Ali Ijaz, and Leo Lahti. 2021. Corpus linguistics and Eighteenth Century Collections Online (ECCO). *Research in Corpus Linguistics*, 9(1):19–34.

Myriam C. Traub, Jacco van Ossenbruggen, and Lynda Hardman. 2015. Impact Analysis of OCR Quality on Research Tasks in Digital Archives. In *Research and Advanced Technology for Digital Libraries*, Lecture Notes in Computer Science, pages 252–263, Cham. Springer International Publishing.

Oxana Vitman, Yevhen Kostiuk, Paul Plachinda, Alisa Zhila, Grigori Sidorov, and Alexander Gelbukh. 2022. Evaluating the Impact of OCR Quality on Short Texts Classification Task. In *Advances in Computational Intelligence*, Lecture Notes in Computer Science, pages 163–177, Cham. Springer Nature Switzerland.

Hadley Wickham. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

A List of features included

agentless passives
amplifiers
analytic negation: not
attributive adjectives
be as main verb
by-passives
causative adverbial subordinators: because
concessive adverbial subordinators: although, though
conditional adverbial subordinators: if, unless
conjuncts
contractions
demonstrative pronouns
direct WH-questions
discourse particles
downtoners
emphatics
existential there
first person plural pronouns
first person singular pronouns
hedges
indefinite pronouns
independent clause coordination
infinitives
necessity modals
nominalizations
other adverbial subordinators
past tense
perfect aspect
phrasal coordination
pied-piping relative clauses
place adverbials
possibility modals
predicative adjectives
predictive modals
present tense
private verbs
pro-verb do
pronoun it
public verbs
second person pronouns
seem & appear
split auxiliaries
split infinitives
suasive verbs

subordinator-that deletion
synthetic negation
that adjective complement
that relative clauses on object position
that relative clauses on subject position
that verb complement
third person personal pronouns
time adverbials
total adverbs
total other nouns
total prepositional phrases
WH relative clauses on object positions
WH relative clauses on subject position
WH-clauses

Automated Generation of Multiple-Choice Cloze Questions for Assessing English Vocabulary Using GPT-turbo 3.5

Qiao Wang
Ralph Rose
Naho Orita
Ayaka Sugawara
Center for English Language Education, Faculty of Science and Engineering
Waseda University

Abstract

A common way of assessing language learners' mastery of vocabulary is via multiple-choice cloze (i.e., fill-in-the-blank) questions. But the creation of test items can be laborious for individual teachers or in large-scale language programs. In this paper, we evaluate a new method for automatically generating these types of questions using large language models (LLM). The VocaTT (vocabulary teaching and training) engine is written in Python and comprises three basic steps: pre-processing target word lists, generating sentences and candidate word options using GPT, and finally selecting suitable word options. To test the efficiency of this system, 60 questions were generated targeting academic words. The generated items were reviewed by expert reviewers who judged the well-formedness of the sentences and word options, adding comments to items judged not well-formed. Results showed a 75% rate of well-formedness for sentences and 66.85% rate for suitable word options. This is a marked improvement over the generator used earlier in our research which did not take advantage of GPT's capabilities. Post-hoc qualitative analysis reveals several points for improvement in future work including cross-referencing part-of-speech tagging, better sentence validation, and improving GPT prompts.

1 Introduction

Vocabulary acquisition lies at the core of foreign language education, forming an essential pillar in most holistic curricula (Alqahtani et al., 2015; Nation, 2022). To measure learners' vocabulary knowledge, a common testing approach is to use multiple-choice cloze questions (hereafter, MCC; Hale et al., 1989): learners see a stem sentence with a blank followed by several one-word options (one correct answer as the "key" plus several distractors that are incorrect answers) and must choose the option which best fills the blank. The following is an example:

> This is a fairly simple process with __ steps.
>
> a. unlimited b. few c. courts d. full

The quality of stems and distractors in MCC questions is crucial. According to previous studies, context clarity and relevance to keys are paramount in stems, meaning that a stem should be free from syntactic errors, with appropriate length to provide context for the key, and should show a good use of the key (Pino et al., 2008). Meanwhile for distractors, part-of-speech (POS) and semantic considerations apply. Specifically, effective distractors should fit syntactically into stems but should be semantically less appropriate than the keys (Brown et al., 2005; Coniam, 1997). Traditionally, the generation of such questions has been a manual endeavor, with pedagogical experts or individual educators crafting content tailored for their classes. Although there is a clear demand for automated tools to facilitate the generation of numerous vocabulary test items, existing applications, as highlighted by studies like Lee et al. (2013), Liu et al. (2005), and Rose (2016, 2020), are either not readily accessible or lack user-friendliness.

Another aspect often neglected in MCC is ensuring that distractors align with students' genuine learning experiences. Students usually engage with vocabulary in structured units or sublists (Schmitt, 1997), which indicates that distractors should be derived only from words they have previously studied. Deviating from this can inadvertently provide clues, allowing students to guess based solely on unfamiliarity, thereby diminishing the test's effectiveness.

In our previous initiative of a web-based vocabulary training and testing application, or "VocaTT"[1], we used the Word Quiz Constructor (WQC, Rose, 2016, 2020) to automatically generate MCC questions for the General Service List and Academic

[1] http://vocatt-server.herokuapp.com/

Word List (Rose et al., 2022). WQC incorporates various features: it tags the POS of input words, crafts a question stem around a chosen keyword from corpus resources, and identifies distractors with matching POS from the input words. It also evaluates the distractors by placing them in the blank and comparing the frequencies in Google books of local tri-grams with that when the keyword is filled. If the former is lower that the latter, then the distractor is considered valid. The modified version of WQC's output has been effectively incorporated into English curricula at a Japanese university and the in-house application "VocaTT". However, it is not without its shortcomings. Most notably, evaluations by human experts flagged quality issues with the generated content. Among the 1128 question stems and 3384 distractors generated for 1128 questions targeting the Academic Word List, the percentages of well-formed stems and distractors were only 34.93% and 38.56%, respectively (Rose et al., 2022).[2] It took the reviewers much effort to manually correct the inappropriate stems and distractors before the questions were imported into the application.

The advent of advanced natural language processing tools, especially models like GPT, provides new opportunities. These models can potentially enhance the quality of automated MCC question generation through the colossal-scale corpora used as their training data and their deep understanding of complex topics (Abdullah et al., 2022). This paper delves into our efforts to create a program that automates MCC question generation incorporating an LLM. We evaluate its effectiveness through human validation and also provide insights into potential refinements, underpinned by a thorough qualitative analysis of both the generation mechanics and the final output.

2 Methodology

Building upon prior work with WQC, the process of automatically generating MCC questions in the program consisted of three distinct phases: preprocessing of input words, stem generation, and distractor selection. A key evolution in the program involves the integration of an LLM during both stem creation and distractor validation. Subsequent sections will delve into the tools employed in this endeavor, followed by a detailed exposition of the program.

2.1 Tools

The program[3] utilized an array of tools, including the Academic Word List (Coxhead, 2000), the GPT-turbo 3.5 API (OpenAI.com) and libraries in Python.

2.1.1 Wordlist

The wordlist at the heart of this research was the Academic Word List (AWL, Coxhead, 2000)[4]. The selection of AWL was driven by its widespread acceptance in academic English courses. Moreover, AWL is divided into 10 sublists, each containing around 60 headwords and their associated word families, which is aligned with the study's premise where students learn vocabulary in smaller sets. Another compelling reason for this choice was our familiarity with the AWL from previous studies with WQC. This past engagement provided a rich dataset that could be leveraged for a direct comparison between the newly developed program and its predecessor. For the aims of this project, only the headwords from the first sublist of the AWL were considered.

2.1.2 LLM API

This study was conducted between May and June of 2023, and we settled on GPT 3.5-turbo as the preferred LLM API. By mid-2023, the performance of GPT 3.5-turbo (hereafter as "GPT") stood out in the domain of LLMs. Its high capabilities in text generation, understanding, and contextual relevance made it an ideal candidate for a project of this nature (Abdullah et al., 2022). In addition, its widespread adoption in the AI community ensured a robust support framework. The active user base often meant quicker solutions to potential issues and a wealth of shared experiences and best practices.

2.1.3 Programming language and libraries

The program was developed using Python. For reading and writing data files, the "pandas" library was utilized. Codes were implemented to interface

[2] Interestingly, this wellformedness rate may suggest the possibility that creating items manually may be more efficient. Although not done in the present work, earlier work with WQC showed that the well-formedness rates of automatically generated items were actually comparable to those for manually-produced items (Rose, 2014, 2016, 2020).

[3] https://github.com/judywq/cloze-generator-with-llm

[4] https://www.wgtn.ac.nz/lals/resources/academicwordlist

with the OpenAI platform using the official OpenAI library[5]. Specifically, GPT was requested to return data in the JSON format, facilitating more straightforward data processing in Python.

Many English words have various POSs and their respective inflected forms, and learners are expected to acquire the most frequent forms and uses of them (Zimmerman, 1997). Take, for instance, the word “account”. Learners should recognize its duality as both a noun and a verb. As a noun, it possesses singular and plural forms, and as a verb, it spans various tenses and forms including the base, present participle, past participle, and third person singular. Given this complexity, an effective library capable of both labeling the POSs of words and extrapolating their inflected forms within each POS became essential. Upon searching for resources using the query “python library for word inflection”, two potential tools were identified: `pyInflect`[6] and `LemmInflect`[7]. After rigorously testing both tools against the AWL words, `LemmInflect` emerged superior in terms of capturing a broader spectrum of word inflections. Also, GPT was able to understand the POS tags from `LemmInflect`. Thus, we adopted `LemmInflect` as the POS and morpheme tagger. The following is an example when tagging the word “distribute”:

```
{'VB': {'distribute'},
 'VBD': {'distributed'},
 'VBG': {'distributing'},
 'VBP': {'distribute'},
 'VBZ': {'distributes'}}
```

2.2 Generation processes

At the outset, we determined the criteria for the question items. With 60 main words in the first sublist, the goal was to design 60 questions that encompassed each of these words, focused on academic English. Drawing from prior research (Graesser, 2001; Brown et al., 2005; Pino et al., 2008; Coniam, 1997), we established guidelines for crafting question stems and choosing distractors. Specifically, the objective was to ensure that question stems remained succinct, not exceeding 20 words in length. Questions were designed to avoid starting with a blank, and any given key should be used only once within a question. Each question would offer three distractors which, while syntactically correct, should be semantically less fitting than the correct word. Figure 1 shows the flow of the generation process and the specific steps will be discussed in detail below.

2.2.1 Pre-processing

For all 60 words, their applicable POSs and inflected forms within each POS were labeled. For the purpose of this study, the combination of a headword and all its inflected forms is defined as a “word group”. The data of word groups was then stored in a csv file for later input.

2.2.2 Question stem generation

In generating a question stem, the program first reads the input file in csv format, randomly selects one word (“key”) from each group, and randomly retrieves one of its POSs. Then, it sends the POS-labeled key to the GPT API and asks it to generate a sentence that shows the use of the key with the specific POS. When GPT returns the sentence, the program replaces the key with a blank to create the question stem.

In writing the prompt, we incorporated the criteria decided earlier, and also provided an example to GPT to ensure better results. The following is an example with the word “creates” tagged “VBZ” (verb, non-3rd person singular present):

```
Generate a sentence with the word "creates" with at
most 20 words. The text domain should be Academic
English. The given word in the sentence has a pos
tag of VBZ. It should not be at the beginning of
the sentence. It should not appear more than once.
Surround it with a backtick.
---
For example, the given word is "account" with
pos tag of "NN". You should yield a sentence in
the following format:
I have an `account` with the bank.
```

Raw response:

```
National income `creates` economic growth and
development in a country.
```

Question stem:

```
National income ____ economic growth and
development in a country.
```

2.2.3 Distractor selection

In deciding the distractors for a question stem, the program first selects 10 words that share the key’s POS from other word groups and stores them as distractor candidates. Then it fills the blank with each of the candidates and asks GPT to judge the syntactic and semantic appropriateness of the candidates in respective complete sentences. If the result for syntactic appropriateness is true and that for semantic appropriateness is false, then the candidate is considered a good distractor. The process

[5] https://github.com/openai/openai-python

[6] https://github.com/bjascob/pyInflect

[7] https://github.com/bjascob/LemmInflect

repeats until the evaluation of the 10th candidate is finished. By then, if the number of good distractors among the 10 candidates is no less than three, the program randomly selects three of them as final distractors for the question stem. Otherwise, it starts another round of distractor selection until three have been found or until the pool of the same POS is depleted. The following shows the prompt and responses in distractor generation and validation, continuing with the previous example:

```
For each of the following words separated by
a comma, when the word is fit into the blank
in the masked sentence, if the syntax of the
sentence is correct yield true for "syntax",
if the semantic meaning of the sentence is
correct yield true for "semantics".
Words: ```sectors, varies, estimates, derives,
processes, functions, legislates, requires,
indicates, assumes```

Masked sentence: ```National income ____
economic growth and development in a country.```
---
Answer in the following JSON structure:
{
  "word 1": {"syntax": true, "semantics": true},
  "word 2": {"syntax": true, "semantics": false}
}
```

Response:

```
{
  "sectors": {"syntax": true, "semantics": false},
  "varies": {"syntax": true, "semantics": true},
 "estimates": {"syntax": true, "semantics": false},
  "derives": {"syntax": true, "semantics": false},
 "processes": {"syntax": true, "semantics": false},
 "functions": {"syntax": true, "semantics": false},
 "legislates": {"syntax": true, "semantics": false},
 "requires": {"syntax": true, "semantics": false},
 "indicates": {"syntax": true, "semantics": true},
  "assumes": {"syntax": true, "semantics": true}
}
```

Good distractors:

```
[sectors<VBZ>, estimates<VBZ>, derives<VBZ>,
processes<VBZ>, functions<VBZ>,
legislates<VBZ>, requires<VBZ>]
```

2.2.4 Output and logging

After a question item is sucessfully generated and stored in the database, the program judges whether the number of question items has met the predefined threshold. If not, the program repeats the generation process. Otherwise, it terminates and returns a csv file containing all question items ("output file") and another csv file containing the prompts and raw responses from GPT ("log file").

3 Data analysis

The research utilized two primary data sources: the output file and the log file, and data analysis was carried out in four steps:

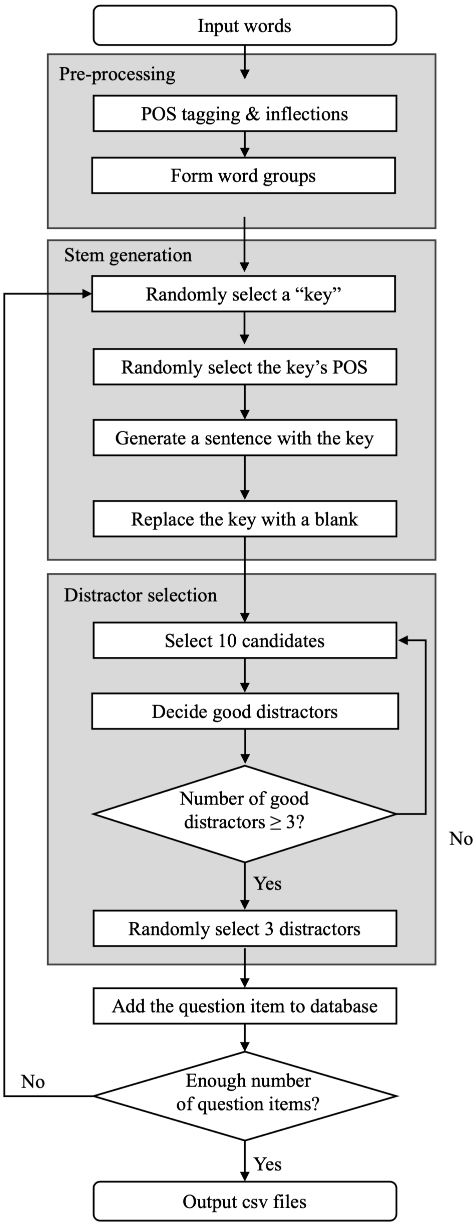

Input words
Pre-processing
POS tagging & inflections
Form word groups
Stem generation
Randomly select a "key"
Randomly select the key's POS
Generate a sentence with the key
Replace the key with a blank
Distractor selection
Select 10 candidates
Decide good distractors
Number of good distractors ≥ 3?
No
Yes
Randomly select 3 distractors
Add the question item to database
No
Enough number of question items?
Yes
Output csv files

Figure 1: Flowchart of the generation process

Step 1. Preliminary output check: Upon receiving the output file, a preliminary examination was initiated. We verified the presence of all question stems, blanks, distractors, and other essential components in the output, ensuring its integrity before progressing to the next phase.

Step 2. Human evaluation: After the preliminary check, two seasoned reviewers were tasked with an independent evaluation of the questions. The reviewers were all English native speakers with more than 20 years of experience teaching academic English at Japanese universities. They had also been involved in a similar project reviewing automatically generated MCC questions on AWL words. Both reviewers underwent training using an evaluation guide. They were asked to judge whether a stem or distractor was appropriate for assessing the vocabulary knowledge of university students, and if not, provide reasons in comment boxes. The criteria for judging appropriateness are as follows:

- **Stem Appropriateness**: a. The context and syntax of an appropriate stem should be understandable even without knowing the key. There should be no grammatical errors. b. An appropriate stem should solicit an accepted use of the key and effectively highlight or emphasize the key.

- **Distractor Appropriateness**: An appropriate distractor should be one that fits grammatically within the stem but is semantically incorrect/remote for the blank. An inappropriate distractor might either be an acceptable answer and/or not fit the stem's syntax.

Once the reviewers' evaluations were submitted, Cohen's *d* and percent agreement were employed to measure inter-rater reliability. In instances where a discrepancy in evaluations arose, a third expert was consulted to deliver the final judgment, substantiated by relevant comments.

Step 3. Human annotation: Subsequent to the human evaluation, items flagged as inappropriate underwent an annotation process by two annotators, who were experienced English teachers with near-native English proficiency working at a Japanese university. Drawing from thematic analysis techniques, the two annotators collaboratively classified the inappropriate stems and distractors, and identified categories and subcategories.

Step 4. Qualitative analysis of the log file: With the annotations in place, an exhaustive qualitative examination was set into motion, leveraging the rich information contained in the log file. The primary objective of this step was to pinpoint the root causes of the identified errors. Such insights are invaluable for refining and enhancing future iterations of the process.

4 Results

Fifteen minutes after its initiation, the program generated the output and log files for analysis. The subsequent sections will detail the results sequentially.

4.1 Preliminary check results

Upon the preliminary check of the output file, two issues were identified: the absence of blanks at key positions in three question stems and two missing distractors in one question item.

Absence of blanks: Three question stems lacked the requisite blanks that should have replaced the keys. In one question stem:

> "*Assessing the validity of the research findings requires a critical and thorough examination.*"

no blank was created at the key "assessing". The absence of the blank can be attributed to the program's case-sensitivity. As the key was placed at the very first of the stem and its first letter was capitalized, it went undetected by the program. The prompt given to GPT explicitly instructed it not to place the keyword at the beginning of the stem. However, in this case, this instruction was disregarded.

In the remaining two cases, the keys were positioned in the middle, not the start, but still no blanks were created. The following is an example:

> *The researchers assumed that the data they collected was reliable and unbiased.*
> Key: assumed

Missing distractors: Two distractors were found missing in one question stem, as follows.

> "*The 'available' resources for research on this topic are limited and need to be expanded.*" Distractor: formula

We created blanks for the affected question stems and flagged the missing distractors as N/A before the output file was sent to the reviewers for evaluation. As a result, the output file contained 60 question items and 178 distractors.

4.2 Annotation results

4.2.1 Stems

The inter-rater reliability for the stems, as assessed by the two reviewers, yielded a Cohen's *d* value of 0.71. This suggests a substantial level of agreement as per (Gisev et al., 2013). Additionally, the agreement rate stood at 0.88. After the third reviewer resolved the disagreements, 15 inappropriate stems were spotted (percentage of appropriate stems = 75%), and subsequently, 15 codes were

Table 1: Error annotation results in stems

Category	Subcategory	Instance
Mechanical	Capitalization	1
Syntax	Determiner	1
	Noun number	1
	Clause conjunction	1
Semantics	Perplexity	1
Key fitness	Rare use/collocation	4
	Syntactic unfitness	6

finalized. The details of these issues are provided in Table 1 and descriptions.

Mechanical issue: The only mechanical issue is related to the technical issue in preliminary check where a capitalized initial letter of the key is expected when the key is placed at the beginning.

Syntax: Some stems contained minor grammatical errors that, though they do not necessarily hinder understanding, should be rectified. Under "Determiner", one stem incorrectly used "various" before the uncountable noun "demographic information". Under "Noun number", "meaning of words" in one stem should have "meaning" in its plural form to match "words". Lastly, under "Clause conjunction", the stem *"I cannot remember the __ for calculating the standard deviation, can you remind me?"* incorrectly linked two clauses with a comma rather than separating them with a period.

Semantics: In the stem *"The study aims to analyze the effectiveness of __ to negative feedback on social media for brand reputation management."*, where the key is "responding", the stem was commented as overly intricate and perplexing without the key.

Key fitness: This category refers to situations where the key was an inappropriate fit in the stem, either syntactically or semantically. Within this, "Rare use/collocation" refers to situations where the key was not the most intuitive or commonly expected answer for the blank. For instance, in the stem *"The research methodology used in this study involved __ a sample of participants through random selection,"* the key, "constituting", might not be the first choice for many. Meanwhile, "Syntactic unfitness" pertains to instances where placing the keys in the stems resulted in subject-verb agreement errors, parts of speech mismatches, or noun number problems. An example is, *"The research project involved testing various __ to determine the most effective strategy,"* where the key "method" doesn't fit syntactically.

4.2.2 Distractors

The inter-rater reliability for the stems, as assessed by the two reviewers, yielded a Cohen's d value of 0.87. This suggests an almost perfect level of agreement as per (Gisev et al., 2013). Additionally, the agreement rate stood at 0.94. A total of 59 inappropriate distractors were identified (percentage of appropriate distractors = 66.85%). During annotation, one of these distractors fell under two subcategories, bringing the instance count to 60. The details of these issues are provided in Table 2 and descriptions.

Table 2: Error annotation results in distractors

Category	Subcategory	Instances
Mechanical	Capitalization	2
Syntax	POS	19
	Verb transitivity	8
	Noun number	3
	Article match	2
	Inflection	1
Semantics	Acceptable answers	24
Others	Similar distractors	1

Mechanical issue: The mechanical issue is also related to capitalization as discussed earlier. The initial letters of the two distractors were not capitalized.

Syntax: In the Syntax category, distractors exhibited grammatical inconsistencies. "POS" mismatches occurred where the distractor's part of speech did not meet the blank's demand, such as instances where an adjective was required, but a noun distractor was chosen. "Verb transitivity" entails errors where some verb distractors didn't fit syntactically within the stems' wider context. Specifically, the key might be an intransitive verb followed by a preposition, but the distractor was a transitive verb incompatible with that preposition. Alternatively, a transitive key verb followed by a noun might have an intransitive distractor. For example, in

> "*The data set __ of various demographic information gathered from the survey participants*,"

while the key "consists" is an intransitive verb aptly followed by "of", the distractor "estimates", a transitive verb, doesn't go syntactically with the preposition "of". Another example is "*It's vital to accurately __ the data to draw meaningful conclusions in research,*" where "interpret" is the suitable transitive verb key, but the distractor "function", being intransitive, doesn't fit following "the data". "Noun number" inconsistencies were noted where distractors were sometimes singular when the context required a plural form. Interestingly, the reverse was not observed. Issues with "Article match" arose, for instance, when the distractor "individual", starting with a vowel sound, was incorrectly preceded by the article "a". Finally, there was a peculiar Inflection case where the Latin inflection "-ae" appeared in the distractor "areae".

Semantics: In such situations, distractors were deemed as acceptable answers by reviewers. For example, in *"The __ of democracy is often discussed in political science classes,"* the key is "policy", but the distractor "environment" was considered an acceptable answer by the reviewers, as in the phrase "environment of democracy".

Others: For "Similar distractors" under this category, the words "labours" and "labour" were both included as separate distractors in the same question item, despite both originating from the root word "labour".

4.3 Log file analysis results

Certain categories or subcategories presented challenges when attempting to identify root causes through the log file. Notably, these encompassed scenarios with missing blanks despite correctly positioned keys during preliminary checks, and acceptable answers in distractors. The latter proved especially prominent, as GPT's interpretation of distractor appropriateness occasionally conflicted with human evaluations, with reviewers viewing such distractors as valid. The underlying reasons for GPT's choices are elusive based on the log file, leading us to hypothesize that the nature of our prompts might be a contributing factor. We'll explore these two unresolved issues further in the limitations section.

The log analysis thus encompassed missing distractors and errors within both stems and distractors. For stems and errors, the anlysis was focused on the category "Syntactic unfitness". This focus was selected given the explicit guidelines on distractor syntactic accuracy and GPT's proven capability in producing syntactically robust sentences; the emergence of such errors was indeed surprising. Recognizing that syntactic errors in distractors often originated from or were influenced by those in stems, the analyses for both were undertaken concurrently. As for other categories and subcategories, they received detailed attention in the annotation results section due to their few occurrences. We'll now present the subsequent analysis results.

4.3.1 Missing distractors

The analysis of the log indicated two potential causes for the issue. Firstly, the key "available" is an adjective, labeled "JJ", and the pool of adjectives was relatively smaller compared to other POSs. There were only enough adjectives to perform one round of distractor selection. Secondly, many distractor candidates seemed to semantically fit the stem, as per the log below:

```
{
  "evident": {"syntax": true, "semantics": true},
  "individual": {"syntax": true, "semantics": true},
  "economy": {"syntax": true, "semantics": true},
  "similar": {"syntax": true, "semantics": true},
  "legal": {"syntax": true, "semantics": true},
  "significant": {"syntax": true, "semantics": true},
  "major": {"syntax": true, "semantics": true},
  "specific": {"syntax": true, "semantics": true},
  "formula": {"syntax": true, "semantics": false},
  "period": {"syntax": true, "semantics": true}
}
```

This indicates that the selection of adjective distractors may need more specific context in the stem to highlight the relevance to the key, which may require lengthening the stems. From the log, POS tagging errors can also be seen. For example, nouns such as "economy", "formula" and "period" are inappropriately labeled adjectives. This point will be discussed in later analysis.

4.3.2 Syntactic unfitness

Three core patterns emerged for causes observed: `LemmInflect`'s assignment of rare or inaccurate POS tags, GPT's alteration of keys, and GPT's misjudgment of syntactic appropriateness upon the integration of distractors into the blanks.

POS tagging errors: `LemmInflect` sometimes mislabeled the POSs of words or assigned rare POS tags. For instance, "period" was mislabeled as "JJ", while "sector" was atypically tagged as "VBP". The tagging errors seem to concentrate in nouns. In particular, this tool displayed a pattern of tagging nouns erroneously as adjectives (e.g., "economy" and "formula" both received JJ tags). Furthermore, it regularly attributed NNS tags to singular nouns, even to uncountable ones like "export". This peculiar behavior implies that countable nouns may be recognized as having two NNS forms: singular and plural forms. Such tagging patterns might explain the use of singular forms when plurals were needed in distractors while the opposite was not observed. Another discovery is that `LemmInflect` includes the Latin inflection "-ae" for many nouns, which led to obsolete words like "areae". These tagging inaccuracies directly led to POS mismatches between distractors and keys, with wrongly tagged distractors getting selected.

Key alterations: In some cases, the tagging errors led to key alterations by GPT based on the incorrect POS. For instance, when `LemmInflect` mislabeled "method" as "NNS" (plural noun), GPT adapted the key to "methods" to match NNS, generating the following sentence:

> *"The research project involved testing various 'methods' to determine the most effective strategy."*

In doing so, when a blank replaced this modified key, the alteration became imperceptible to reviewers, who thus judged that the original key "method" would not fit syntactically into the stem.

However, not all key alterations by GPT were justified. There were cases where despite accurate tagging by `LemmInflect`, GPT replaced the key or its POS. In one case, GPT substituted the key, "major". While the key was correctly labeled "VBP", as seen in "major in", GPT replaced it with an entirely different word: "indicate", though with the same POS of "VBP". The resulting sentence is as follows:

> *"The results of the study 'indicate' the need for further research on the topic."*

When a blank was created, it became evident to the reviewers that "major" did not align syntactically with the blank in the stem.

The alteration of the POS of a key also led to errors in the syntax of the stem and the POS of distractors. In one case, `LemmInflect` appropriately tagged "labour" as "VBP", but GPT altered it to "VBZ" and chose a third-person singular noun as the subject, causing a grammatical error pertaining to subject-verb agreement in the stem when the VBP key was filled. Another example involves the key "finances", correctly tagged as VBZ. However, GPT generated a sentence using "finances" as an NNS: *"The professor emphasizes that understanding one's 'finances' is an important life skill."* In this context, the distractor "indicates", which would have been appropriate if "finances" was used as VBZ, becomes misaligned since the sentence now requires an NNS.

Misjudgement of distractor's syntactic fitness: Despite these mismatches, GPT frequently certified the syntactic appropriateness of distractors. In certain scenarios, this could be attributed to the language model's broad definition of syntactic validity, such as treating two-noun combinations like "bus station" as syntactically correct when an adjective distractor was needed. Yet, in other cases, GPT simply overlooked the errors.

Errors sometimes resulted from a combination of tagging errors, key alterations and misjudgement of fitness. An example involved "sector" being tagged as "VBP", which led to the selection of all VBP distractors, including "involve". When GPT adjusted this to "NNS", it formed: "The government 'sectors' that are responsible for public health need more funding." Here, GPT failed to spot the syntactic incongruence when incorporating "involve" as a distractor. Table 3 shows the attribution of errors in stems and distractors.

Table 3: Summary of error attribution in stems and distractors

Error Attribution	Number
`LemmInflect`	19
GPT	21
Both (`LemmInflect` and GPT)	7

5 Conclusions

Upon comparing this program with the prior Word Quiz Constructor, a marked improvement was evident in the generation of accurate question stems (75% as compared to 34.93%) and distractors (66.85% as compared to 38.56%). Despite this progress, manual correction is still needed before such questions can be imported into the application. It may be better to iterate the program to improve its accuracy rather than asking experts to correct the questions. Human validation highlighted areas for refinement. Foremost among these is the accuracy of POS tagging–a pivotal component in ensuring relevant question stems and well-formed distractors. Besides, higher accuracy of syntactic and semantic judgment outcomes from GPT is necessary, in which better prompts and iterative nature of the GPT API may play a crucial role.

Looking ahead, if these improvements are effectively implemented, the program has the potential to be transformed into a web-based application. We plan to integrate the program into the current "VocaTT" application to enable teachers and students to upload their custom word lists and set variables for question generation. In this way, without any coding expertise, they can receive ready-to-use question items.

Limitations

While our current method showed significant improvement over the older Word Quiz Constructor (WQC) in generating MCC question items, it is not without limitations.

Central among these is the lack of validation of POS and inflection results from `LemmInflect`, which led to many inappropriately tagged keys and distractors. In the pre-processing phase, it would be judicious to incorporate a cross-validation step by comparing POS tags identified by `LemmInflect` with another reliable source. This source could be another Python-based POS tagger or even GPT itself and only POSs and inflected forms recognized by both sources should be adopted in the word groups.

Another limitation is the lack of stem validation, which resulted in missing blanks, incorrect key placements, and unintentional key changes. The validation can be done by by prompting GPT to check for the key's presence and position in the sentence and ensuring it retains the chosen POS. It would also be beneficial to leverage GPT to review complete sentences for syntactic or semantic issues, which can help reduce grammatical and collocation errors. On a related note, the 20-word limit in stems may have contrained the context in highlighting a key in some cases. By extending these stems to encompass, say, approximately 30 words, it could foster a richer context and thus bolster the relevance of the key within the stem.

The third limitation lies in distractor validation. The present emphasis of the prompt lies on the individual distractors, leading GPT to misjudge their syntactic and semantic appropriateness in quite a few cases. Instead, examining the full sentences with the distractors inserted could be more telling. This would not only identify issues like verb transitivity but also check the overall coherence of the sentences, providing a comprehensive assessment of the distractors' fit.

Another notable limitation concerns the small sample size and the program's speed. Some potential issues may have gone unnoticed as only 60 question items were generated. Despite the small sample size, the generation process took 15 minutes, with the majority of this duration dedicated to calling the GPT API and waiting for its response. The introduction of stem validation as suggested earlier will necessitate additional prompts, potentially lengthening the wait time. Finding a solution to expedite this process will be an ongoing challenge.

A further limitation relates to the intended audience of the generated MCC items. In the present work, reviewers were instructed to evaluate the items assuming they would be presented to university student learners of English as a second/foreign language. Naturally, the results could look quite different if the intended audience were, say, high school students or adult learners in the community. A full-fledged generation system would need to account for this at the level of GPT interactions or through filtering mechanisms.

Ethics Statement

Although the reviewers in this study were paid for their review work, there is no conflict of interest and they are independent from the institution the researchers are affiliated with. Data were analyzed impartially, and the results presented are an honest representation of the research findings.

Acknowledgements

This paper is a part of research outcomes funded by Waseda University Grant for Special Research Project (project number: 2023R-020). The researchers would like the thank the reviewers and annotators for their work in the study.

References

Malak Abdullah, Alia Madain, and Yaser Jararweh. 2022. Chatgpt: Fundamentals, applications and social impacts. In *2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–8. IEEE.

Mofareh Alqahtani et al. 2015. The importance of vocabulary in language learning and how to be taught. *International journal of teaching and education*, 3(3):21–34.

Jonathan Brown, Gwen Frishkoff, and Maxine Eskenazi. 2005. Automatic question generation for vocabulary assessment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 819–826.

David Coniam. 1997. A computerised english language proofing cloze program. *Computer Assisted Language Learning*, 10(1):83–97.

Averil Coxhead. 2000. A new academic word list. *TESOL quarterly*, 34(2):213–238.

Natasa Gisev, J Simon Bell, and Timothy F Chen. 2013. Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, 9(3):330–338.

Arthur C Graesser. 2001. *Question generation as a learning multiplier in distributed learning environments*, volume 1121. US Army Research Institute for the Behavioral and Social Sciences.

Gordon A. Hale, Charles W. Stansfield, Donald A. Rock, Marilyn M. Hicks, Frances A. Butler, and JR John W. Oller. 1989. The relation of multiple-choice cloze items to the test of english as a foreign language. *Language Testing*, 6(1):47–76.

K. Lee, S. Kweon, H. Kim, and G. Lee. 2013. Filtering-based automatic cloze test generation. In *Proceedings of Speech and Language Technology in Education (SLaTE)*, page 72–76.

C. Liu, C. Wang, Z. Gao, and S. Huang. 2005. Applications of lexical information for algorithmically composing multiple-choice cloze items. In *Proceedings of the 2nd Workshop on Building Educational Applications Using NLP*, page 1–8.

Paul Nation. 2022. Teaching and learning vocabulary. In Eli Hinkel, editor, *Handbook of Practical Second Language Teaching and Learning*, pages 397–408. Routledge, New York, NY, USA.

Juan Pino, Michael Heilman, and Maxine Eskenazi. 2008. A selection strategy to improve cloze question quality. In *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains. 9th International Conference on Intelligent Tutoring Systems, Montreal, Canada*, pages 22–32.

Ralph L. Rose. 2014. Automated vocabulary quiz creation using online and offline corpora. Poster presentation at Teaching and Language Corpora (TaLC) conference, Lancaster University, UK.

Ralph L. Rose. 2016. Automatic word quiz construction using regular and simple english wikipedia. In *INTED2016 Proceedings*, pages 8032–8040. IATED.

Ralph L. Rose. 2020. Improving the production efficiency and well-formedness of automatically-generated multiple-choice cloze vocabulary questions. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7094–7101.

Ralph L. Rose, Naho Orita, Ayaka Sugawara, and Qiao Wang. 2022. Evaluation dataset of multiple-choice cloze items for vocabulary training and testing. In *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers*, pages 292–298.

Norbert Schmitt. 1997. Vocabulary learning strategies. In Norbert Schmitt and Michael McCarthy, editors, *Vocabulary: Description, acquisition and pedagogy*, pages 199–227. Cambridge University Press, Cambridge, UK.

Cheryl Boyd Zimmerman. 1997. Do reading and interactive vocabulary instruction make a difference? an empirical study. *TESOL quarterly*, 31(1):121–140.

Explicit References to Social Values in Fairy Tales: A Comparison between Three European Cultures

Alba Morollon Diaz-Faes
Institute for the Study of
Literature and Tradition
NOVA University of Lisbon
1069-061 Lisbon, Portugal
albafaes@fcsh.unl.pt

Carla Sofia Ribeiro Murteira
Institute for the Study of
Literature and Tradition
NOVA University of Lisbon
1069-061 Lisbon, Portugal
School of Communication
and Media Studies, IPL
Portugal
carlamurteira@fcsh.unl.pt

Martin Ruskov
University of Milan
20123 Milan, Italy
martin.ruskov@unimi.it

Abstract

The study of social values in fairy tales opens the possibility to learn about the communication of values across space and time. We propose to study the communication of values in fairy tales from Portugal, Italy and Germany using a technique called word embedding with a compass to quantify vocabulary differences and commonalities. We study how these three national traditions of fairy tales differ in their explicit references to values. To do this, we specify a list of value-charged tokens, consider their word stems and analyse the distance between these in a bespoke pre-trained Word2Vec model. We triangulate and critically discuss the validity of the resulting hypotheses emerging from this quantitative model. Our claim is that this is a reusable and reproducible method for the study of the values explicitly referenced in historical corpora. Finally, our preliminary findings hint at a shared cultural understanding and the expression of values such as Benevolence, Conformity, and Universalism across European societies, suggesting the existence of a pan-European cultural memory.

1 Introduction

Culture is defined "as a common heritage of a set of beliefs, norms, and values" (US DHHS 2001), that influences an individual's cognition and behaviour (Wong, 2013). Social values are understood as standards or criteria of the desirable, thus they guide the selection or evaluation of behaviours, policies, people, and events (Schwartz et al., 2020). Building on this understanding of values as a cornerstone of culture, we turn to literature as a mirror reflecting these values across different cultural contexts in the past. Developments in NLP, in particular word embeddings, have allowed for the quantitative analysis of historical corpora (Miaschi and Dell'Orletta, 2020; Rodriguez and Spirling, 2022).

With this work we want to test the limits of an approach for studying the social values present in fairy tales, one of the most widely spread forms of popular narratives and a privileged genre for the identification of patterns of cultural exchange, as they have historically migrated across different cultures and periods, creating a rich tapestry of storytelling traditions. In particular, we study the aggregated explicit tokens mapped on the values proposed by the Theory of Basic Human Values (Schwartz, 1992, 2012) across fairy tale corpora from three different European traditions – namely Portugal, Italy and Germany – in order to compare the quantitative representations and analyse the emerging patterns. We do this by first finding the stemmed matches of these tokens and enriching the text with the corresponding annotation. After that we employ a word embedding with a compass (Di Carlo et al., 2019) and clique percolations (Palla et al., 2005) to highlight the semantic variation between the three national corpora.

A critical investigation of the results of our method finds that its results correspond to findings of previous research. We also find indications that despite the differences on the expression of values in the three compared countries, it seems that Values of Benevolence (quality of interpersonal relationships), Conformity (respect for social norms and expectations) and Universalism (protection of the welfare of people and nature) have remained consistent in fairy tales across the three national traditions, which we also view as confirmation of the validity of our approach for the study of values embedded in historical, literary corpora.

2 Background

The study of explicit references of values in fairy tales is related to the accumulated social attitudes up to the historical period of codification of the

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 62–75
December 1-3, 2023 ©2023 Association for Computational Linguistics

tales. To our knowledge, no systematic research of this wide topic exists. As such, we view it as being at the crossroads between the socio-historical, literary study of fairy tales, and the psychological study of social values which is shaped by contemporary research. On the other hand, such a study at scale would not be possible without the instruments and methods of computational humanities and word embeddings in particular.

2.1 Unpacking Fairy-Tale Studies from the Brothers Grimm to Digital Humanities

The late 18th century witnessed the rise of folklore studies as part of a quest for national and cultural identity, particularly in Europe (Schacker, 2003). Jakob and Wilhelm Grimm, riding the tide of renewed interest in popular culture among the upper-class intelligentsia, became pivotal figures in this domain. They first published their fairy-tale collection *Children's and Household Tales* in 1812, striving to present a pure German narrative tradition, untouched by foreign influence, particularly the French (Teverson, 2013). This publication sparked what would become the 19th century's golden age of fairy tales across Europe. This was a time of growing urbanisation, industrialization, and literacy. Scholars and nationalists, fearful of losing invaluable oral traditions due to these rapid societal changes, began the collection and preservation of folklore (Ostry, 2013). Among these custodians were collectors and writers such as Italy's Giuseppe Pitré and Portugal's Consiglieri Pedroso, whose texts feature prominently here alongside the Grimms'. Their work, heavily inspired by the Grimms, was driven by a desire to distil and dialectically construct their nations' cultural legacy.

Despite the nationalistic intentions of Brothers Grimm and others who embarked on preserving what they thought to be distinct national narratives, the study of fairy tales reveals as much about the interconnectedness of cultures as it does about their uniqueness. Fairy tales, at their core, are a blend of narratives that "migrate on soft feet" (Warner and Warner, 2016), indicating that they traverse and interweave across generic, geographical and temporal boundaries, sometimes in untraceable ways. Thus, while the Grimms and others sought to capture and enshrine a uniquely national heritage, their work also serves to underscore the similarities between narrative traditions.

Unpicking these similarities and differences, however, can prove to be quite a complex task. As scholars are frequently dependent on translations, the potential for misinterpretation or loss of nuanced meanings during this process is high. Translations, like the ones by Margaret Hunt, Thomas Crane and Henriqueta Monteiro used here, are enormously valuable artefacts, but must be recognised as acts of literary adaptation that might differ from the originals (Haase, 2016). Further complicating matters, the comparative analysis of several national traditions involves processing vast quantities of text to identify patterns. This challenge extends beyond the study of fairy tales and into the comparative study of literature as a whole.

In response to these challenges, digital humanities and computer-assisted literary studies offer innovative methodologies. Computational methods, in particular, aid in identifying and assessing literary patterns across scales, from individual texts to entire fields and systems of cultural production (Wilkens, 2015). These new approaches, to which our work is a contribution, help produce new types of evidence that enrich and expand humanities research. Indeed, computational approaches to fairy tales have already successfully been deployed in studies such as "Computational analysis of the body in European fairy tales" (Weingart and Jorgensen, 2013). In this study, the authors used digital humanities research methods to analyse the representations of gendered bodies in European fairy tales. They created a manually curated database listing every reference to a body or body part in a selection of 233 fairy tales, and its analysis revealed that the gender and age of fairy-tale protagonists correlate in ways that indicate societal biases, particularly against the ageing female body. A further exploration of gender bias in fairy tales is presented in "Are Fairy Tales Fair?" (Isaza et al., 2023). This study employs computational analysis to dissect the sequence of events in fairy tales, revealing that one in four event types exhibit gender bias when not considering temporal order, and that female characters are more likely to experience gender-biased events at the start of their narrative arcs. These studies underscore the potential of distant reading, data analysis and visualisation as powerful tools in the comparative study of fairy tales, particularly when used alongside subject expert close reading (Moretti, 2022). Nevertheless, perceptions and attitudes towards gender represent just a fraction of the broader societal values spectrum.

2.2 The expression of values across space and time in European Fairy Tales

Values are regarded as a shared societal understanding of what constitutes *good*, *wrong*, *fair*, *unfair*, *just*, *right* or *ethical* behaviour (Haidt, 2013; Kesebir and Haidt, 2010; Turiel, 2005). Values are cognitive representations of an individual's biological needs, an individual's requirements in interpersonal coordination, and the institutional demands focused on group welfare and survival (Schwartz and Bilsky, 1987). Nonetheless, it is crucial to acknowledge the significance of cultural and individual influence in the development and expression of values. Cultural Psychology postulates that human behaviours result from the reciprocal interaction between cultural and individual psyche (Shweder, 1991; Cohen, 2011; Schwartz et al., 2020). However, the manifestation of behaviors and values is contingent upon context and situation, implying that similar cultural processes might serve or facilitate different purposes based on cultural context (Rogoff, 2003; Schwartz et al., 2020). Therefore, one could examine variations in the expression of values across different regions and periods through the analysis of historical corpora. This stems from the expectation that literature can be used as a vehicle for the expression of cultural norms and values, thereby reflecting the distinct ideological attributes of the writers and the regions from which it emerges (Albrecht, 1956). Several Theories have been proposed to summarise values across different cultures (for a review of theories see Ellemers et al (2019)). In this paper we focus on the Theory of Basic Values (Schwartz, 2012), since it found validity expression across several cultures (e.g. (Spini, 2003; Schwartz et al., 2001, 2014; Davidov et al., 2008), and it has been applied in the study of European values (e.g., European Social Survey). The Theory of Basic Human Values (Schwartz, 2012) comprises 10 human values that are fuelled by four different and opposite motivations: Openness to Change vs. Conservation, Self-Transcendence vs. Self-Enhancement as observed in Figure 1.

Openness to Change relates to an individual's need for independence of thought, action, and feelings, and readiness for change, therefore comprises the values of Self-direction, Stimulation, and partly Hedonism. On the other hand, Conservation relates to the values of Security, Conformity and Tradition, as it emphasises the individual's needs for order, preservation of the past, and resistance to change. Self-enhancement considers the individual's needs to pursue their own interests, success, and dominance over others, therefore comprises the values of Power, Achievement, and partly Hedonism. On the other hand, Self-transcendence considers the values of Universalism and Benevolence, to focus on the welfare and better interests of others. For a definition of specific values, see Table 1.

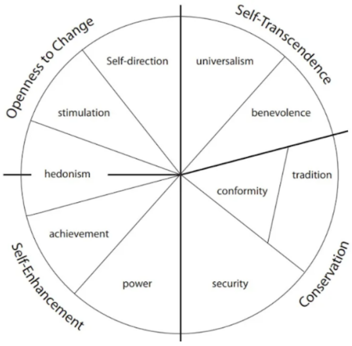


Figure 1: Theoretical model of relations among ten motivational types of values (Schwartz, 2012).

Europeans can be regarded as having a common identity (Castano, 2004) that is expressed through their way of life, values and culture, and that has been building since ancient times (Pagden, 2002; Pinheiro et al., 2012) leading to the establishment of a broad set of European Values. Values such as human dignity, freedom, democracy, equality, rule of law, and human rights have been declared as the values of the European Union, to form "a society in which inclusion, tolerance, justice, solidarity and non-discrimination prevail" (EU, 2020). Based on several empirical studies and policy making guidelines, these values correspond to Schwartz's values of Universalism, Self-Direction, and Benevolence (for more information see: (Scharfbillig et al., 2021; Murteira, 2024). If these values are presumed to have been shared to some degree across the European territory since antiquity, it stands to reason that they could have been variously conveyed through fairy tales across the three regions under analysis. Constructs such as values can either be assessed by explicit or implicit measures. A psychological construct is implicitly assessed when the individual "is unaware that a psychological measurement is taking place, this type of measure is often used to assess values, attitudes, stereotypes, and emotions in social cognition research" (APA, 2023). On the other hand, a psychological construct is explicitly assessed when the "individual

Table 1: The definition of each of the ten motivational types of values (Schwartz, 2012).

Value	Definition
Security	Safety, harmony, and stability of society, of relationships, and of self.
Tradition	Respect, commitment, and acceptance of the customs and ideas that traditional culture or religion provide the self. Maintaining and preserving cultural, family or religious traditions.
Conformity	Restraint of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms.
Self-Direction	Independent thought and action-choosing, creating, exploring.
Stimulation	Excitement, novelty, and challenge in life.
Hedonism	Pleasure and sensuous gratification for oneself.
Achievement	Personal success through demonstrating competence according to social standards.
Power	Social status and prestige, control or dominance over people and resources.
Benevolence	Preservation and enhancement of the welfare of people with whom one is in frequent personal contact.
Universalism	Understanding, appreciation, tolerance, and protection of the welfare of all people and of nature.

is aware that a psychological measurement is taking place" (APA, 2023). Putting it simply, values can be measured explicitly when individuals are directly asked about values, and implicitly when the individuals are not aware of the measurement, because values are assessed using indirect questioning methods. Bearing in mind that art is a behavioural expression of culture that serves several purposes, including the *form of order*, which is the need for psychological and mental organisation of experiences (Dissanayake, 1980), we can hold the reasonable expectation that the historical corpora under analysis will reflect, to a degree, the explicit and implicit cultural ways and behaviours of societies in which these fairy tales were written. The presence of these values in our corpora was assessed by vocabulary quantification techniques through the development of Word Embeddings that communicate values in fairy tales.

2.3 Using Word Embeddings to Quantify Vocabulary Differences

Word embeddings have emerged as an important instrument for the quantitative analysis of textual corpora. These are mappings of vocabulary onto a multidimensional numerical space, based on their occurrences (Mikolov et al., 2013; Rodriguez and Spirling, 2022). Different techniques for creating word embeddings exist, but their common general principle is "a word is characterised by the company it keeps". It is useful to distinguishing between two categories of word embeddings: i) static (also called type-based) - those that feature a single representation for a word token, and ii) contextual (also called token-based) - those that distinguish between different representations of each word to capture potential differences in meanings, according to the surrounding context (Miaschi and Dell'Orletta, 2020; Lenci et al., 2022). Whereas contextual word embeddings better capture the richness of vocabulary, static word embeddings perform better on smaller corpora which do not have the volume that would allow for the semantic richness necessary for multiple meanings (Ehrmanntraut et al., 2021). Arguably, this is due to the fact that in a small thematic corpus, typically meanings are restricted by the context of its compilation.

A widespread approach that allows to overcome the challenge of small corpora and their lack of richness, is the combination of pre-training with a huge[1] generic corpus and the subsequent fine-tuning with the corpus of interest. However, corpora of these huge dimensions are inevitably contemporarily written, and due to cultural and linguistic change over time inevitably introduce unwanted biases. In confirmation of this consideration, in their particular context Manjavacas and Fonteyn (2022) observed that training from the ground up is more effective than fine-tuning of pre-existing models.

When it comes to comparing the word embeddings representing different corpora, a widespread approach is the so-called semantic change detection (Tahmasebi et al., 2021). Since for intercultural comparison, "change" might wrongly suggest a transition from one culture to the other, in the context here the phrase "semantic variation" would be more accurate. Still, whenever techniques for semantic change detection do not rely on any particular diachronic properties of the underlying corpora, we claim they could be reused also for synchronic linguistic analysis. More specifically we claim that

[1] Some widely established contextual models like BERT are trained on a corpus that includes the entire contents of Wikipedia which comprises of 2.5 billion word tokens (Devlin et al., 2019), others use training sets that are many orders of magnitude larger (Dodge et al., 2021)

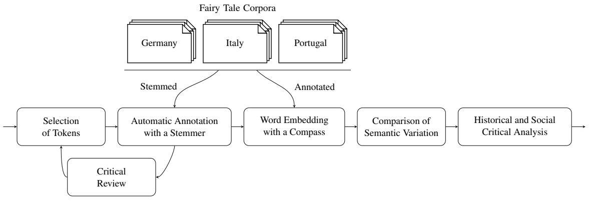

Fairy Tale Corpora
Germany
Italy
Portugal
Stemmed
Annotated
Selection of Tokens
Automatic Annotation with a Stemmer
Word Embedding with a Compass
Comparison of Semantic Variation
Historical and Social Critical Analysis
Critical Review

Figure 2: The outline of the process we followed.

an approach called temporal word embedding with a compass (Di Carlo et al., 2019) is applicable, for culture-specific rather than time-specific distinctive corpora. This approach consists of first creating an embedding on a cumulative corpus containing all texts from the different cultures to be considered. Then, from this baseline (compass) word embedding, further fine-tuning is performed on each of the corpora, to be compared so as to create culture-specific word embeddings. The result is a different (numerical) vector representations of each particular word token, which allows for quantitative comparisons between them, as done previously (Ferrara et al., 2022; Di Carlo et al., 2019).

3 Method

Our study of the explicit references to values in fairy tales follows the process illustrated in Figure 2. To provide an outline, it starts with the identification of tokens that represent values of interest. We group these tokens in groups that we consider to be synonyms. Then, we automatically annotate all occurrences in the text of the stems representing the considered tokens. Once this is done, we manually analyse the produced annotations to identify ambiguities and mistakes in our identification of tokens. The purpose of this analysis is to better understand the semantics behind their occurrences, in order to refine the selection of tokens and identify potential tokens representing multiple values. Finally, we apply a static word embedding with a compass and perform critical analysis on the differences from the resulting vector spaces.

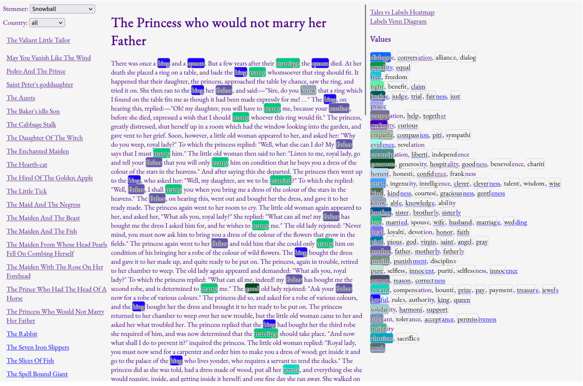
Fairy Tales Corpora. The corpus selection had several stages. First we focused on the Grimms' *Children's and Household Tales*, using Margaret Hunt's 1884 English translation. We manually selected 30 tales that span well-known and beloved stories and lesser known ones, so as to provide a comprehensive representation of the entire collection. Then we selected 30 Portuguese and 30 Italian tales taken from two important contemporary collections to the Grimms': *Portuguese Folktales* by Consiglieri Pedroso, translated to English in 1882 by Henriqueta Monteiro; and *Italian Popular Tales*, collected and translated to English in 1885 by Thomas Frederick Crane. These collections were chosen due to their cultural significance and their temporal proximity to the Grimms' collection, aiming to offer a comparative perspective on 19th century fairy tales across different European cultures.

Selection of Tokens. Assuming that the historical corpora are themselves mirrors of social behaviours and ways of living in societies in which the fairy tales were written, we are interested in the explicit expressions of values in the texts. Starting from Schwartz's model and the European core values, we initially compose a list of tokens that represent these values, based on three empirical studies regarding value-specific tokens. This list of tokens contains words that were selected from two dictionary studies about values, where each word is associated with a specific Schwartz's value. (Schwartz, 1992; Lindeman and Verkasalo, 2005; Murteira, 2024). For instance, the token "Peace" is associated with the value of Universalism, and the token "Cooperation" is associated with the value of Benevolence (see Table 4 in Appendix). Then we perform automatic identification of explicit references of values. We do this using stemming (Jabbar et al., 2020) on both the token lists and the fairy tale texts. This is because, in contrast e.g. to lemmatisation, stemming reduces different word forms the same originating token. We use the Snowball stemmer algorithm (Porter, 2001) to identify all occurrences of the stemmed tokens in the corpora and tag (i.e. annotate) them with a label corresponding to the group of synonym tokens.

Critical Review. We then critically analyse and refine by adapting tokens according to the desired annotation. This was done using a specifically de-

Table 2: Quantitative descriptors of the corpora. When we refer to tokens, we mean the ones that were identified by our automated annotation process. Complete list of included texts is available in Table 3 in Appendix.

Corpus	Texts	Symbols	Words	Tokens
Germany	30	306 475	59 500	1840
Italy	30	234 158	45 223	1808
Portugal	30	231 149	44 887	1439



Stemmer: Snowball
Country: all

The Valiant Little Tailor
May You Vanish Like The Wind
Pedro And The Prince
Saint Peter's goddaughter
The Aunts
The Baker's idle Son
The Cabbage Stalk
The Daughter Of The Witch
The Enchanted Maiden
The Hearth-cat
The Hind Of The Golden Apple
The Little Tick
The Maid And The Negress
The Maiden And The Beast
The Maiden And The Fish
The Maiden From Whose Head Pearls Fell On Combing Herself
The Maiden With The Rose On Her Forehead
The Prince Who Had The Head Of A Horse
The Princess Who Would Not Marry Her Father
The Rabbit
The Seven Iron Slippers
The Slices Of Fish
The Spell Bound Giant

The Princess who would not marry her Father

There was once a king and a queen. But a few years after their marriage the queen died. At her death she placed a ring on a table, and bade the king marry whomsoever that ring should fit. It happened that their daughter, the princess, approached the table by chance, saw the ring, and tried it on. She then ran to the king her father, and said:—"Sire, do you know that a ring which I found on the table fits me as though it had been made expressly for me! . . " The king, on hearing this, replied:—"Oh! my daughter, you will have to marry me, because your mother, before she died, expressed a wish that I should marry whoever this ring would fit." The princess, greatly distressed, shut herself up in a room which had the window looking into the garden, and gave vent to her grief. Soon, however, a little old woman appeared to her, and asked her: "Why do you weep, royal lady?" To which the princess replied: "Well, what else can I do? My father says that I must marry him." The little old woman then said to her: "Listen to me, royal lady, go and tell your father that you will only marry him on condition that he buys you a dress of the colour of the stars in the heavens." And after saying this she departed. The princess then went up to the king, who asked her: "Well, my daughter, are we to be married?" To which she replied: "Well, father, I shall marry you when you bring me a dress of the colour of the stars in the heavens." The father, on hearing this, went out and bought her the dress, and gave it to her ready made. The princess again went to her room to cry. The little old woman again appeared to her, and asked her, "What ails you, royal lady?" She replied: "What can ail me! my father has bought me the dress I asked him for, and he wishes to marry me." The old lady rejoined: "Never mind, you must now ask him to bring you a dress of the colour of the flowers that grow in the fields." The princess again went to her father and told him that she could only marry him on condition of his bringing her a robe of the colour of wild flowers. The king bought the dress and gave it to her made up, and quite ready to be put on. The princess, again in trouble, retired to her chamber to weep. The old lady again appeared and demanded: "What ails you, royal lady?" To which the princess replied: "What can ail me, indeed! my father has bought me the second robe, and is determined to marry me." The good old lady rejoined: "Ask your father now for a robe of various colours." The princess did so, and asked for a robe of various colours, and the king bought her the dress and brought it to her ready to be put on. The princess returned to her chamber to weep over her new trouble, but the little old woman came to her and asked her what troubled her. The princess replied that the king had bought her the third robe she required of him, and was now determined that the marriage should take place. "And now what shall I do to prevent it?" inquired the princess. The little old woman replied: "Royal lady, you must now send for a carpenter and order him to make you a dress of wood; get inside it and go to the palace of the king who lives yonder, who requires a servant to tend the ducks." The princess did as she was told, had a dress made of wood, put all her jewels, and everything else she would require, inside, and getting inside it herself; and one fine day she ran away. She walked on

Tales vs Labels Heatmap
Labels Venn Diagram

Values

dialogue, conversation, alliance, dialog
equality, equal
free, freedom
right, benefit, claim
justice, judge, trial, fairness, just
peace
cooperation, help, together
curiosity, curious
empathi, compassion, piti, sympathi
evidence, revelation
emancipation, liberti, independence
generous, generosity, hospitality, goodness, benevolence, chariti
honest, honesti, confidence, frankness
smart, ingenuity, intelligence, clever, cleverness, talent, wisdom, wise
kind, kindness, courtesi, graciousness, gentleness
know, able, knowledge, ability
brother, sister, brotherly, sisterly
love, married, spouse, wife, husband, marriage, wedding
loyal, loyalti, devotion, honor, faith
pieti, pious, god, virgin, saint, angel, pray
mother, father, motherly, fatherly
punish, punishment, discipline
pure, selfless, innocent, puriti, selflessness, innocence
correct, reason, correctness
reward, compensation, bounti, prize, pay, payment, treasure, jewels
lawful, rules, authority, king, queen
solidarity, harmoni, support
tolerant, tolerance, acceptance, permissiveness
majority
altruism, sacrifice
truth

Figure 3: Screenshot of a browsing page from the bespoke web instrument that reviews the produced annotations. Another view shows a clickable heatmap as Figure 7 in Appendix, which allows for a distant reading view.

veloped for the purpose web interface (Figure 3, accessible online at `https://tales.ko64eto.com`) that allows for a review of the texts in the corpora with the results of the automatic annotation highlighted in different colours. The outcome of this was a series of decisions to adapt the token selection as a way to refine it and guide subsequent iterations of this annotation process. Correspondingly, following this approach inspired by grounded theory (Rieger, 2019), the ultimately proposed list of tokens in this study emerges from exploration of the corpus and is not a result of deductive hypothesis research. We provide a statistical overview of the results of this annotation process in Table 2 and in Appendix we provide both the complete final version of our tokens and a Venn diagram of the occurrences of groups of synonym token across the three corpora.

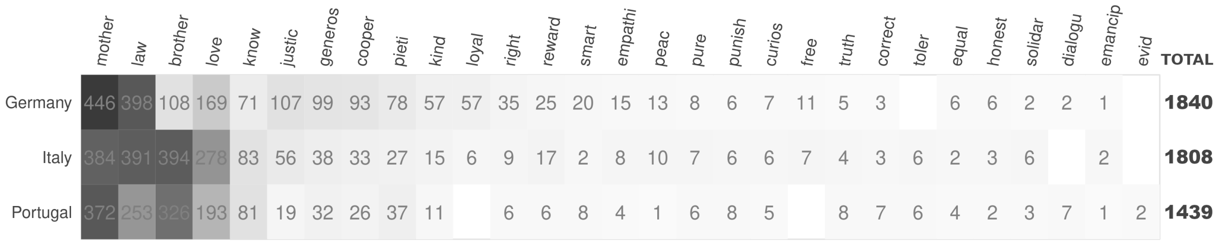
Word Embedding with a Compass. Due to the historical nature of the corpora we study and in order to avoid contaminating them with external biases from pre-training, we organise our analysis following the word embedding with a compass approach (Di Carlo et al., 2019). To do this, we create one generic culture-agnostic shared embedding from scratch containing all three corpora. Then, starting from this compass, we independently create three parallel fine-tunings for each of the cultures. For the creation of the compass, to avoid the possible introduction of biases, we chose not to include any further possible texts, neither from any of our three contexts, nor from others. Our approach to syntactic identification of references to values, is not contextual, i.e. we treat a reference to a value-related stemmed token as the same for all its identified uses. In our critical review step we examine the validity of this generalisation. To represent the annotations in the word embedding algorithm, before and after each identified occurrence of a token we insert an indication of the corresponding group of synonym tokens (i.e. the first token in it).

Comparison of Semantic Variation. The word embedding allows measuring contextual similarity between words, thus speaking of "change" and "variation". Once we have the three word embeddings for the cultural corpora, for each of them we consider only the distances between groups of tokens (represented by the annotation label, i.e. the first token in each synonym group) and experimentally define a similarity threshold above which we consider a pair of tokens to have a relating edge between them in a graph representation of tokens) in order to use clique percolations clustering with k=2 (Palla et al., 2005). In other words, for all similarities above that threshold we consider the corresponding tokens to be related in that embedding, and distances above the threshold mean the corresponding tokens are not. This results in a clustering that might assign one token to multiple clusters. It might also bring two tokens into the same cluster even if the distance between them is greater than the threshold, as long as there is a "bridge" of other tokens in between to connect them.

Historical and Social Critical Analysis. At the end of our method we analyse the quantitative results using critical analysis from the perspectives of both literary studies and psychological research. This allows us to cross-validate (e.g. through triangulation (Noble and Heale, 2019)) our results with the established body of research and thus get an indication of their theoretical validity.

4 Results

An important part of the results of our approach is the reflective inspection of the produced automated annotation and possible corrections for these. An overall conclusion of this process is that, expectedly, the most impactful tokens capture the values they were intended to match well. The most important token that did not correspond to our initial interpretation was "faith". We originally ascribed the label "faith" to the value of "piety," indicating religious devotion. However, a careful examination of our corpus revealed an intriguing trend. The term



	mother	law	brother	love	know	justic	generos	cooper	pieti	kind	loyal	right	reward	smart	empathi	peac	pure	punish	curios	free	truth	correct	toler	equal	honest	solidar	dialogu	emancip	evid	TOTAL
Germany	446	398	108	169	71	107	99	93	78	57	57	35	25	20	15	13	8	6	7	11	5	3		6	6	2	2	1		**1840**
Italy	384	391	394	278	83	56	38	33	27	15	6	9	17	2	8	10	7	6	6	7	4	3	6	2	3	6		2		**1808**
Portugal	372	253	326	193	81	19	32	26	37	11		6	6	8	4	1	6	8	5		8	7	6	4	2	3	7	1	2	**1439**

Figure 4: Frequencies of identified occurrences of tokens across the three corpora. A more detailed heatmap between texts and labels is available on Figure 7 in Appendix.

"faith," contrary to our initial classification, typically expressed affiliation with "loyalty," mainly as per the usage patterns in various Grimm tales, particularly in "Faithful Johannes" (a German tale). As a consequence, we ascribe the token "faith" to the value associated to "loyalty".

Another token that provides an interesting example is "father," due to its potential multiple associations. On the one hand, it could represent "caring," similar to "mother," but on the other, it could be a symbol of authority (Hopp et al., 2021). When exploring the corpora, we found that "father" was predominantly associated with "caring," with a remarkable exception in "The Maiden and the Fish" (Portugal), where one out of four instances appeared associated with authoritative power.

A third, less impactful token we considered was "patient," which was initially associated with "kindness." However, an analysis of the corpus found that its usage related exclusively to an individual receiving medical treatment, and we consequently excluded it from our analysis.

Figure 4 shows the references to values by countries, according to the ascribed tokens. A more detailed mapping of occurrences of tokens in particular texts is provided in Figure 7 in the Appendix. From the resulting comparison of clusters across corpora, noteworthy is the one defined around tokens related to "mother." As the Venn diagram on Figure 5 shows, while in our German and Portuguese corpora it appears together with "brother", in the Italian and Portuguese corpora it also appears in relation to "know." Only in Germany does it relate to "generous."

4.1 Historical Analysis

Dolores Buttry elucidates on the usage of "faith" in Grimm tales to exclusively mean "loyalty," and not "piety." She writes that the related values of faithfulness and loyalty (which are "Treu" and "Treue" in German) have been foundational virtues in Germany since ancient times (Buttry, 2011). Stories such as "Faithful Johannes," but also "The Frog King," exemplify extreme loyalty towards superiors, illustrating the importance of fidelity and respect for authority in their various manifestations. Buttry characterises the tale of the faithful/loyal servant as an enduring archetype, highlighting the recurring appearance of the words "Treu" (faithful) and "Treue" (loyalty, fidelity) in German tales (Buttry, 2011). She further suggests that, while respect for authority and the sanctity of oaths were nearly universal concepts before these stories were collected, they seem to have retained their vitality and cultural significance particularly in German-speaking traditions. This idea finds further support in one of the only non-German occurrences of "faith" in our corpus, as the label appears in "The Story Of Catherine and Her Fate," a Sicilian tale first collected by Swiss-German folklorist Laura Gozenbach.

It is also interesting to examine how values manifest in tales from different cultural contexts. In our results, we found that values of "piety" and "empathy" appeared clustered together in Italian and Portuguese tales, but not in German ones. This may be explained by the different religious traditions in all three countries, since both Italy and Portugal were majoritarily Catholic regions at the time the tales were collected, while there was a strong Protestant presence in the German territory. Indeed, Jack Zipes (2002) writes that the Grimms' tales portrayed the main values of Protestant ethics

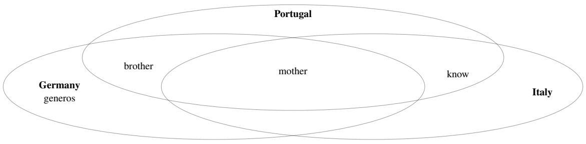


Figure 5: An illustration of the degree of overlap across the three national corpora for the token "mother"

and the bourgeois enlightenment. The heroes in their tales are predominantly concerned with self-preservation and the acquisition of wealth, and they assist others, including animals, only when they perceive a potential gain for themselves, demonstrating a calculated approach to empathy and compassion. This model of behavior, Zipes argues, exemplifies the general Protestant ethic of the time, and so empathy, although occasionally appearing in the Grimms' tales, is not a dominant theme (Zipes, 2002). We may advance the possibility that the differing religious ethos of Italy and Portugal would place more emphasis on empathy as it relates to Catholic piety.

4.2 Social Analysis

Frequency analysis shows that tokens such as "mother," "law," brother," and "love" have a strong presence (more than 100 appearances) across the three countries under analysis. Based on the elaborated correspondence between tokens and the Theory of Basic Values (see Appendix), the words "mother," "brother" and "love" are connected to Benevolence, and "law" is connected to Conformity. In Germany, the token "justice" has also a strong presence, and is connected with the value of Universalism which stands for the protection and welfare of all people and nature. Considering that the value Benevolence stands for the good quality of social connections between people, and Conformity stands for the preservation of social/cultural expectations and norms, then we could infer that these tales describe several social dynamics. The tales' plots are representative of social dynamics among fictional characters that may resemble society, in order to describe the quality of human relationships and social/cultural norms in place.

Interestingly, some differences across countries are expressed by the tokens' frequency related to Benevolence, Conformity and Universalism. For instance, in Germany, "mother" seems to be a stronger reference for communication of Benevolence than "brother" when compared to Portugal and Italy. Also, "love" seems to be a stronger reference for communication of Benevolence in Italy than in Germany and Portugal. However, in Germany, we may note that tokens such as "generous" and "cooperation" reinforce the communication and expression of Benevolence in those tales. When it concerns the need for rules and social welfare, it seems that in Germany and Italy the token "law" is frequently used when compared to Portugal to express the value of Conformity. Finally, Germany shows a strong presence of token "justice" in their tales, which highlights the importance of Universalism in these tales and the need to convey the respect for human rights and dignity. In sum, while Portugal, Italy and Germany communicate strongly the values of Benevolence and Conformity, it seems that Germany also communicates the value of Universalism. Despite the differences between countries, it seems that European Values of Benevolence and Universalism are being communicated by the tales across countries.

5 Discussion and Conclusion

While the proposed approach is still in its infancy, and the emerging results would require more thorough examination, our preliminary analysis provides some concrete evidence that European Values have been a long-standing element in European cultural communication through fairy tales. The corpus analysis across different cultures revealed a significant variety in the representation of values. For example, the affiliation of the token ”faith” with ”loyalty” rather than ”piety,” particularly in German culture, illustrates the role of cultural and historical contexts in shaping value representations. Similarly, the differential clustering of ”piety” and ”empathy” in Italian and Portuguese tales compared to German tales further underscores the influence of religious and socio-cultural contexts in value representation. Interestingly, despite these differences, the analysis revealed a strong commonality across all three cultures, pointing at the communication of European Values through tales. Tokens associated with Benevolence, Conformity, and Universalism manifested frequently across fairy tales of all three countries. This finding is particularly noteworthy because it suggests a shared cultural understanding and expression of these values across European literary production, and, possibly and by extension, across European societies, thus hinting at the existence of a pan-European cultural memory.

We have identified clear limitations in our approach. Working at the syntactic level, both in terms of stemming and static word embeddings, limits the possibility to capture nuances, and with this some noise is introduced in the analysis. However, contrary to our expectations, our detailed analysis revealed that ambiguities are only isolated cases. This is valid to the extent that in none of these cases a token bore semantic ambiguity that was a dichotomy rather than an outlier so that it could undermine the general results.

The focus on explicit references, unsurprisingly, resulted in an inability to annotate tokens such as “democracy” in the tales, as they were only implicitly referenced. Therefore, exploring methods to apply semantic word embeddings to historical texts could be a potential way to address not just explicit, but also implicit references to values (Ferrara et al., 2023). While such approaches already exist (Montanelli and Periti, 2023), we believe further attention should be paid to the possibility that the pre-trained embeddings may introduce biases unrelated to the corpus under study.

This work provides a foundational understanding of how European Values are represented in literary texts and highlights the potential of computational linguistics in cultural studies. This study encourages further interdisciplinary research in the field of literary studies, cultural analytics, and computational linguistics to expand our understanding of cultural values and their historical evolution.

Acknowledgments

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101004949. This document reflects only the author’s view and the European Commission is not responsible for any use that may be made of the information it contains.

References

Milton C. Albrecht. 1956. Does Literature Reflect Common Values? *American Sociological Review*, 21(6):722–729.

APA, 2023. 2023. APA Dictionary of Psychology.

Dolores Buttry. 2011. Treue in three tales by the Brothers Grimm. *Forum for World Literature Studies*, 3(2):166–173.

Emanuele Castano. 2004. *European identity: A social-psychological perspective*, Governance in Europe, chapter 3. Rowman & Littlefield, Lanham, MD.

Dov Cohen. 2011. Cultural Psychology. Technical report, Oxford University Press. Type: dataset.

Henriqueta Monteiro Consiglieri Pedroso. 1882. *Portuguese Folk-Tales*. WikiSource.

Thomas Frederick Crane. 2017. *Italian Popular Tales*. Project Gutenberg.

E. Davidov, P. Schmidt, and S. H. Schwartz. 2008. Bringing Values Back In: The Adequacy of the European Social Survey to Measure Values in 20 Countries. *Public Opinion Quarterly*, 72(3):420–445.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Valerio Di Carlo, Federico Bianchi, and Matteo Palmonari. 2019. Training Temporal Word Embeddings with a Compass. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6326–6334.

Ellen Dissanayake. 1980. Art as a Human Behavior: Toward an Ethological View of Art. *The Journal of Aesthetics and Art Criticism*, 38(4):397.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. ArXiv:2104.08758 [cs].

Anton Ehrmanntraut, Thora Hagen, Leonard Konle, and Fotis Jannidis. 2021. Type- and token-based word embeddings in the digital humanities. In *Proceedings of the Conference on Computational Humanities Research 2021*, number 2989 in CEUR Workshop Proceedings, pages 16–38, Aachen.

Naomi Ellemers, Jojanneke Van Der Toorn, Yavor Paunov, and Thed Van Leeuwen. 2019. The Psychology of Morality: A Review and Analysis of Empirical Studies Published From 1940 Through 2017. *Personality and Social Psychology Review*, 23(4):332–366.

EU, 2020. 2020. The EU values. Last accessed 30 July 2023.

Alfio Ferrara, Stefano Montanelli, and Martin Ruskov. 2022. Detecting the semantic shift of values in cultural heritage document collections (short paper). In *Proceedings of the 1st Workshop on Artificial Intelligence for Cultural Heritage*, number 3286 in CEUR Workshop Proceedings, pages 35–43, Aachen.

Alfio Ferrara, Sergio Picascia, and Elisabetta Rocchetti. 2023. Augustine of Hippo at SemEval-2023 Task 4: An Explainable Knowledge Extraction Method to Identify Human Values in Arguments with SuperASKE. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1044–1053, Toronto, Canada. Association for Computational Linguistics.

J. Grimm, W. Grimm, and M. Hunt. 1884. *Household Tales by the Brothers Grimm*. George Bell and Sons.

Donald Haase. 2016. Challenges of Folktale and Fairy-Tale Studies in the Twenty-First Century. *Fabula*, 57(1-2):73–85.

Jonathan Haidt. 2013. *The righteous mind: why good people are divided by politics and religion*, 1st vintage books ed edition. Vintage Books, New York. OCLC: 900283765.

Frederic R. Hopp, Jacob T. Fisher, Devin Cornell, Richard Huskey, and René Weber. 2021. The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior Research Methods*, 53(1):232–246.

Paulina Toro Isaza, Guangxuan Xu, Akintoye Oloko, Yufang Hou, Nanyun Peng, and Dakuo Wang. 2023. Are Fairy Tales Fair? Analyzing Gender Bias in Temporal Narrative Event Chains of Children's Fairy Tales.

Abdul Jabbar, Sajid Iqbal, Manzoor Ilahi Tamimy, Shafiq Hussain, and Adnan Akhunzada. 2020. Empirical evaluation and study of text stemming algorithms. *Artificial Intelligence Review*, 53(8):5559–5588.

Selin Kesebir and Jonathan Haidt. 2010. Morality (in Handbook of Social Psychology).

Alessandro Lenci, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani. 2022. A comparative evaluation and analysis of three generations of Distributional Semantic Models. *Language Resources and Evaluation*, 56(4):1269–1313.

Marjaana Lindeman and Markku Verkasalo. 2005. Measuring Values With the Short Schwartz's Value Survey. *Journal of Personality Assessment*, 85(2):170–178.

Enrique Manjavacas and Lauren Fonteyn. 2022. Adapting vs. Pre-training Language Models for Historical Languages. *Journal of Data Mining & Digital Humanities*, NLP4DH(Digital humanities in...):9152.

Alessio Miaschi and Felice Dell'Orletta. 2020. Contextual and Non-Contextual Word Embeddings: an in-depth Linguistic Investigation. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 110–119, Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space.

Stefano Montanelli and Francesco Periti. 2023. A Survey on Contextualised Semantic Shift Detection. ArXiv:2304.01666 [cs].

Franco Moretti. 2022. *Falso movimento: la svolta quantitativa nello studio della letteratura*. Extrema ratio. Nottetempo, Milano.

Carla Murteira. 2024. Towards an ontology of values: Elaboration of a dictionary of values' related words based on the theory of social values from schwartz. In preparation.

Helen Noble and Roberta Heale. 2019. Triangulation in research, with examples. *Evidence-Based Nursing*, 22(3):67–68.

Elaine Ostry. 2013. *Social Dreaming*, 0 edition. Routledge.

Anthony Pagden, editor. 2002. *The idea of Europe: from antiquity to the European Union*. Woodrow Wilson Center series. Woodrow Wilson Center Press ; Cambridge University Press, Washington, DC : Cambridge ; New York.

Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. 2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818.

Teresa Pinheiro, Beata Cieszynska, and José Eduardo Franco, editors. 2012. *Ideas of | for Europe*. Peter Lang D.

M.F. Porter. 2001. Snowball: A language for stemming algorithms. Accessed 27 July 2023.

Kendra L. Rieger. 2019. Discriminating among grounded theory approaches. *Nursing Inquiry*, 26(1):e12261.

Pedro L. Rodriguez and Arthur Spirling. 2022. Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research. *The Journal of Politics*, 84(1):101–115.

Barbara Rogoff. 2003. *The cultural nature of human development*. Oxford University Press, Oxford [UK] ;a New York.

Jennifer Schacker. 2003. *National Dreams: The Remaking of Fairy Tales in Nineteenth-Century England*. University of Pennsylvania Press.

M. Scharfbillig, L. Smillie, D. Mair, M. Sienkiewicz, J. Keimer, R. Pinho Dos Santos, H. Vinagreiro Alves, E. Vecchione, and Scheunemann L. 2021. *Values and identities: a policymaker's guide*. Publications Office of the European Union, Luxembourg. OCLC: 1289305231.

Seth J. Schwartz, Ágnes Szabó, Alan Meca, Colleen Ward, Charles R. Martinez, Cory L. Cobb, Verónica Benet-Martínez, Jennifer B. Unger, and Nadina Pantea. 2020. The Convergence Between Cultural Psychology and Developmental Science: Acculturation as an Exemplar. *Frontiers in Psychology*, 11:887.

Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pages 1–65. Elsevier.

Shalom H. Schwartz. 2012. An Overview of the Schwartz Theory of Basic Values. *Online Readings in Psychology and Culture*, 2(1).

Shalom H. Schwartz and Wolfgang Bilsky. 1987. Toward a universal psychological structure of human values. *Journal of Personality and Social Psychology*, 53(3):550–562.

Shalom H. Schwartz, Gian Vittorio Caprara, Michele Vecchione, Paul Bain, Gabriel Bianchi, Maria Giovanna Caprara, Jan Cieciuch, Hasan Kirmanoglu, Cem Baslevent, Jan-Erik Lönnqvist, Catalin Mamali, Jorge Manzi, Vassilis Pavlopoulos, Tetyana Posnova, Harald Schoen, Jo Silvester, Carmen Tabernero, Claudio Torres, Markku Verkasalo, Eva Vondráková, Christian Welzel, and Zbigniew Zaleski. 2014. Basic Personal Values Underlie and Give Coherence to Political Values: A Cross National Study in 15 Countries. *Political Behavior*, 36(4):899–930.

Shalom H. Schwartz, Gila Melech, Arielle Lehmann, Steven Burgess, Mari Harris, and Vicki Owens. 2001. Extending the Cross-Cultural Validity of the Theory of Basic Human Values with a Different Method of Measurement. *Journal of Cross-Cultural Psychology*, 32(5):519–542.

Richard A. Shweder. 1991. *Thinking through cultures: expeditions in cultural psychology*. Harvard University Press, Cambridge, Mass.

Dario Spini. 2003. Measurement Equivalence Of 10 Value Types From The Schwartz Value Survey Across 21 Countries. *Journal of Cross-Cultural Psychology*, 34(1):3–23.

Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen. 2021. *Computational approaches to semantic change*. Language Science Press, Berlin.

Andrew Teverson. 2013. *Fairy Tale*, 0 edition. Routledge.

Elliot Turiel. 2005. Thought, Emotions, and Social Interactional Processes in Moral Development. In *Handbook of Moral Development*, pages 7–35. Psychology Press.

US DHHS 2001. 2001. *Mental health: culture, race, and ethnicity: a supplement to Mental health, a report of the Surgeon General*. Report of the Surgeon General Series. Department of Health and Human Services, U. S. Public Health Service, Washington, D.C. OCLC: 47911898.

Marina Warner and Marina Warner. 2016. *Once Upon a Time: A Short History of Fairy Tale*. Oxford University Press, Oxford, New York.

S. Weingart and J. Jorgensen. 2013. Computational analysis of the body in European fairy tales. *Literary and Linguistic Computing*, 28(3):404–416.

Matthew Wilkens. 2015. Digital Humanities and Its Application in the Study of Literature and Culture. *Comparative Literature*, 67(1):11–20.

Paul T. P. Wong. 2013. Positive Psychology. In Kenneth D Keith, editor, *The Encyclopedia of Cross-Cultural Psychology*, 1 edition, pages 1021–1027. Wiley.

Jack Zipes. 2002. *The Brothers Grimm: From Enchanted Forests to the Modern World*. Palgrave Macmillan US, New York.

A Appendix

The source code of the annotation analysis tool is available at:

`https://github.com/umilISLab/moreever/`.

Continued with figures and tables on the following pages.

Table 3: The Fairy Tales included in the corpora. The Italian corpus includes several collectors. When not indicated, collected by Giuseppe Pitré. Otherwise, 1. Vittorio Imbriani; 2. Domenico Comparetti; 3. Laura Gozenbach; and 4. Carolina Coronedi-Berti.

Germany (Grimm et al., 1884)	Italy (Crane, 2017)	Portugal (Consiglieri Pedroso, 1882)
• Allerleirauh • Briar Rose • Cinderella • Faithful John • Fitcher's Bird • Frau Trude • Godfather Death • Hansel And Grethel • King Thrushbeard • Little Red Cap • Little Snow White • Old Sultan • One Eye Two Eyes And Three Eyes • Our Lady's Child • Rapunzel • Rumpelstiltskin • Snow White And Rose Red • Strong Hans • The Frog King Or Iron Henry • The Giant And The Tailor • The Girl Without Hands • The Jew Among Thorns • The Juniper Tree • The King Of The Golden Mountain • The Lazy Spinner • The Robber Bridegroom • The Six Servants • The Three Spinners • The Two Kings Children • The Valiant Little Tailor	• Brother Giovannone • Cinderella[3] • Don Firriulieddu • Godmother Fox • King Bean Giuseppe Bernoni • Little Chick Pea Tuscan variant • Pitidda • Snow White Fire Red • The Cat And The Mouse • The Cistern • The Cloud[3] • The Crumb In The Beard[3] • The Crystal Casket • The Dancing Water The Singing Apple And The Speaking Bird • The Doctor's Apprentice • The Fair Angiola[3] • The Fair Fiorita[3] • The King Of Love • The King Who Wanted A Beautiful Wife[3] • The Lord St Peter And The Apostles • The Parrot Which Tells Three Stories • The Sexton's Nose • The Shepherd Who Made The King's Daughter Laugh[3] • The Stepmother • The Story Of Catherine And Her Fate[3] • The Story Of Crivoliu[3] • The Story Of St James Of Galicia[3] • The Three Admonitions • Thirteenth • Water And Salt	• May You Vanish Like The Wind • Pedro And The Prince • Saint Peter's Goddaughter • The Aunts • The Baker's Idle Son • The Cabbage Stalk • The Daughter Of The Witch • The Enchanted Maiden • The Hearth-cat • The Hind Of The Golden Apple • The Little Tick • The Maid And The Negress • The Maiden And The Beast • The Maiden And The Fish • The Maiden From Whose Head Pearls Fell On Combing Herself • The Maiden With The Rose On Her Forehead • The Prince Who Had The Head Of A Horse • The Princess Who Would Not Marry Her Father • The Rabbit • The Seven Iron Slippers • The Slices Of Fish • The Spell Bound Giant • The Spider • The Step Mother • The Three Citrons Of Love • The Three Little Blue Stones • The Three Princes And The Maiden • The Tower Of Ill Luck • The Two Children And The Witch • The Vain Queen

Table 4: List of tokens mapped with the values proposed in the Theory of Basic Values from Schwartz (Schwartz, 1992; Lindeman and Verkasalo, 2005; Murteira, 2024).

Token	Synonyms	Value
dialogu	conversation	Universalism
equality	equality, equal	Universalism
free	free	Self-Direction
right	right, claim	Universalism
justic	justice, judge, trial, fairness, just	Universalism
peace	peace	Universalism
cooper	help, together	Benevolence
curios	curiosity, curious	Self-Direction
empathi	compassion, pity	Conformity
evid	evidence	Universalism
emancip	liberty	Self-Direction
generous	hospitality, goodness	Benevolence
honest	honest, confidence	Benevolence
smart	clever, cleverness, wise	Achievement
kind	kind, kindness, graciousness, gentleness	Conformity
know	know, able, knowledge	Achievement
brother	brother, sister, brotherly, sisterly	Benevolence
love	love, married, wife, husband, marriage, wedding	Benevolence
loyal	honor, faith	Benevolence
pieti	piety, pious, god, virgin, saint, angel, pray	Tradition
mother	mother, father, motherly, fatherly	Benevolence
punish	punish, punishment	Conformity
pure	pure, innocent, innocence	Tradition
correct	correct, reason, correctness	Universalism
reward	reward, prize, pay, treasure, jewels	Power
law	lawful, king, queen	Power
solidar	harmony, support	Benevolence
toler	acceptance, permissiveness	Universalism
truth	truth	Universalism

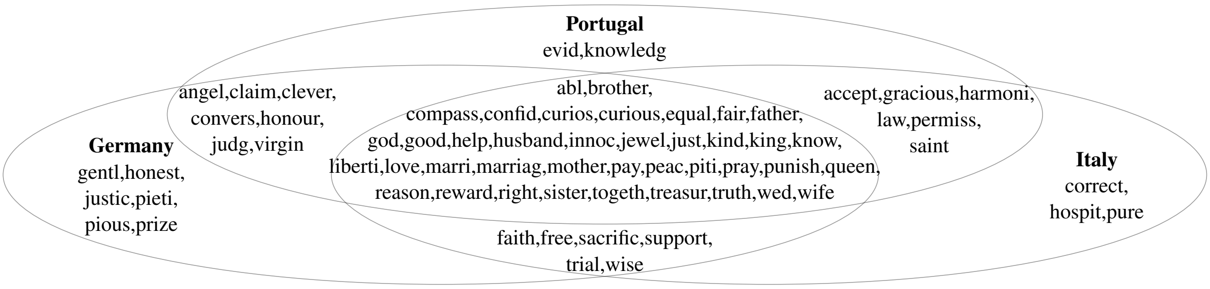

Portugal
evid,knowledg
angel,claim,clever,
convers,honour,
judg,virgin
abl,brother,
compass,confid,curios,curious,equal,fair,father,
god,good,help,husband,innoc,jewel,just,kind,king,know,
liberti,love,marri,marriag,mother,pay,peac,piti,pray,punish,queen,
reason,reward,right,sister,togeth,treasur,truth,wed,wife
accept,gracious,harmoni,
law,permiss,
saint
Germany
gentl,honest,
justic,pieti,
pious,prize
Italy
correct,
hospit,pure
faith,free,sacrific,support,
trial,wise

Figure 6: A Venn diagram showing the occurrences of tokens across the national corpora.

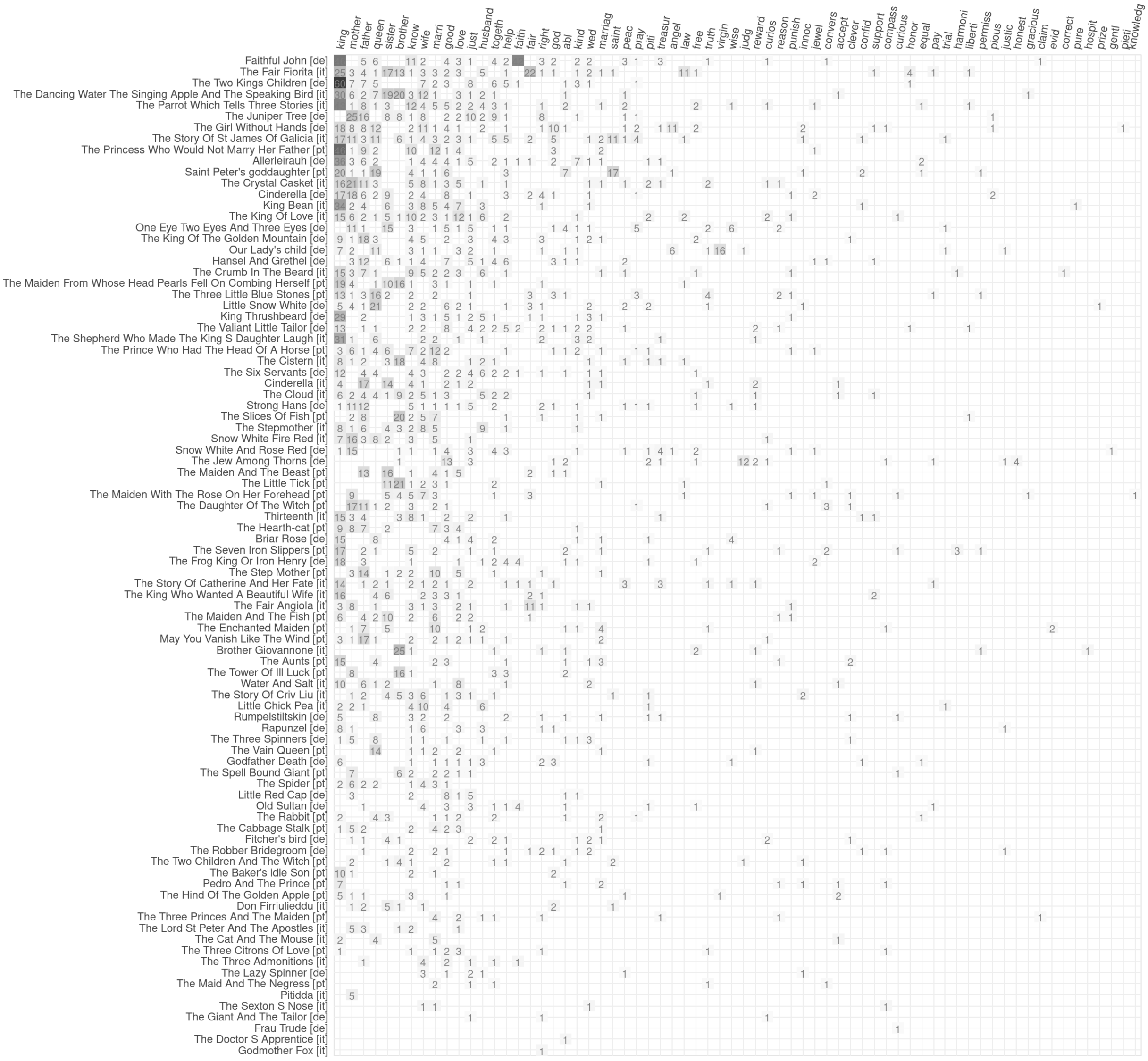


Figure 7: Counts of identified occurrences of tokens across the texts of the three corpora. An interactive version of this heatmap is available at https://tales.ko64eto.com. In it clicking on a number takes you to the corresponding text for easier review.

The Stylometry of Maoism: Quantifying the Language of Mao Zedong

Maciej Kurzynski
Lingnan University, Hong Kong
maciej.kurzynski@ln.edu.hk

Abstract

Recent advances in computational stylometry have enabled scholars to detect authorial signals with a high degree of precision, but the focus on accuracy comes at the expense of explainability: powerful black-box models are often of little use to traditional humanistic disciplines. With this in mind, we have conducted stylometric experiments on Maospeak, a language style shaped by the writings and speeches of Mao Zedong. We measure per-token perplexity across different GPT models, compute Kullback–Leibler divergences between local and global vocabulary distributions, and train a TF-IDF classifier to examine how the modern Chinese language has been transformed to convey the tenets of Maoist doctrine. We offer a computational interpretation of ideology as reduction in perplexity and increase in systematicity of language use.

1 Introduction

Stylometry, the quantitative analysis of literary style, has been extensively used to study various authors' writing styles, leveraging linguistic features such as word frequencies, sentence length, and syntactic patterns (Stamatatos, 2009; Neal et al., 2017). Historically, stylometry dates back to analyses of narrative style in Shakespeare (Burrows, 1987) and attempts to identify the authors of the disputed Federalist Papers (Mosteller and Wallace, 1963; Tweedie et al., 1996). The advent of computational methodologies has significantly enhanced the scope and depth of stylometric analyses, allowing for the examination of larger textual corpora and the incorporation of multifaceted linguistic features, such as lexical richness, syntactic complexity, and semantic coherence (Seroussi et al., 2014; Sari et al., 2018). Computational stylometry has evolved to include a range of techniques, from Principal Component Analysis (PCA) to various machine learning algorithms and large language models (LLMs), enhancing the reliability of stylistic differentiation (Ruder et al., 2016; Ou et al., 2023).

However, the focus on precision metrics in authorship attribution comes at the expense of interpretability and applicability in humanistic research and teaching. The fact that a particular author uses definite articles or certain grammatical particles more often than others might shed light on their subconscious stylistic preferences, and it might even be decisive in distinguishing their stylistic fingerprints, but similar analyses can hardly explain why a given language has been particularly successful at conveying political messages and furthering domestic mobilization, as was the case of Maoism in post-1949 China.

2 Quantifying Maospeak

Maospeak, Mao-style prose, or *maowenti* (毛文体, also called *maoyu* 毛语), is a set of stylistic features influenced by the writings and speeches of Mao Zedong (1893-1976), the leader of the Chinese communist revolution (Li, 1998). Maospeak has had a transformative impact on the way people in China express themselves, and it continues to affect the everyday language in the PRC even today. Consider the following examples:

- 共产党内部也有斗争。不斗争就不能进步, 不和平。八亿人口, 不斗行吗?

  There is also [class] struggle within the Communist Party. Without struggle, there can be no progress, no peace. With a population of 800 million, how can we not struggle?

- 在我们的面前有两类社会矛盾，这就是敌我之间的矛盾和人民内部的矛盾。这是性质完全不同的两类矛盾。

  There are two types of social contradictions in front of us: the contradictions between ourselves and our enemies and the contradictions within the people. These are two types of contradictions with completely different natures.

Such repetitive, redundant, and depersonalized sentences are a staple of the *Little Red Book*, a compilation of Mao's quotations and a condensed example of the Maoist prose. Despite its thematic coherence, however, Maospeak is not simply a set of LDA topics: revolutionary themes appear in Marx, Lenin, Stalin, and Mao, for example, but their writing styles are recognizably different. Neither is it a matter of function words, since Maospeak exudes an affective strength and a clarity of purpose that cannot be reduced to

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 76–81
December 1-3, 2023 ©2023 Association for Computational Linguistics

Corpus	Number of Tokens (Words)	Vocabulary Size	Total Characters	Type-Token Ratio
Maospeak	1,684,294	55,113	2,890,605	0.0327
Contemporary	18,219,475	515,468	28,509,185	0.0283
Eileen Chang	994,025	70,484	1,532,465	0.0710
Mo Yan	2,546,989	131,317	3,978,079	0.0516

Table 1: Dataset Statistics

the most frequent grammatical particles alone. While thoroughly researched by scholars in humanistic disciplines (Ji, 2003; Link, 2013; Schoenhals, 2007) and vividly debated on public platforms (Sun, 2012; Link, 2012; Laughlin, 2012; Barmé, 2012), so far there has been very little computational engagement with the language of Mao Zedong (Huang and Shi, 2022). Maospeak poses an interesting challenge to modern stylometry and remains to this day a controversial issue, especially given the diverse opinions surrounding writers like Mo Yan (b. 1955), the 2012 Nobel Prize laureate who has been accused by critics of inheriting the Maoist style in his descriptions of war-time violence and brutality (Link, 2012; Sun, 2012).

2.1 Dataset

Our dataset (Table 1) consists of four corpora: the selected works of Mao Zedong, collected novels and short stories of Mo Yan and Eileen Chang (Zhang Ailing), and a larger compilation of 102 novels published by 62 writers active in the post-Mao era. In this study, we treat Mao Zedong's writings as a proxy for Maospeak, as it was chiefly through quotations from Mao that the discourse of class struggle and popular militarization spread across the PRC, thus shaping the everyday language. This influence was particularly evident during the Cultural Revolution (1966-1976), when inability to quote the *Little Red Book* could be taken as proof of reactionary politics (Ji, 2003, 151). We chose Eileen Chang (1920-1995) as a control writer because she spent most of her life outside mainland China, and her writings arguably lack communist influences; the mixture of contemporary Chinese writing serves as another control, offering a sample of modern literary Chinese. We preprocessed Mao Zedong's writings by removing footnotes and lines of text shorter than 50 characters to filter out titles, dates, and signatures frequently attached by editors to his letters and communiques. Since Chinese does not use spaces between words, all texts have been segmented with the `spaCy` parser for Chinese `zh_core_web_lg`.[1]

2.2 Perplexity

One way in which different literary styles can be compared is by evaluating the perplexity of their representative texts. A low perplexity indicates that the text is more predictable (Kilgarriff and Rose, 1998; Józefowicz et al., 2016). Auto-regressive language models such as GPT are especially useful in this regard: in the pre-training phase, the model iterates over a large amount of data and learns to predict the next token given a sequence of tokens; these learned predictions can be then used to calculate the "surprisingness" of the actual words encountered in a text.

The formula to calculate the average per-token perplexity for a corpus C consisting of M texts, each containing K words, is represented as follows:

$$\mathcal{P}(C) = \exp\left(-\frac{1}{T}\sum_{j=1}^{M}\sum_{i=1}^{K}\log p(w_{ji}|w_{j,1:i-1})\right) \tag{1}$$

In (1), $p(w_{ji}|w_{j,1:i-1})$ represents the probability of the i-th word in the j-th text of the corpus, given the preceding words in that text. The equation computes the average of log probabilities across all non-special tokens T. If the model assigns high probabilities to the actual words, the average log probability will be less negative, which, after negation and exponentiation, will lead to a lower perplexity. Conversely, lower probabilities for the actual words will result in a higher perplexity.

In this experiment, we used three publicly available Chinese GPT-2 models: **Wenzhong 2.0**,[2] **uer-gpt2**,[3] and **gpt2-base-chinese** from **CKIP Lab**.[4] Using multiple models allowed us to not only compare the results but also mitigate the impact of the pretraining data: some models might have "seen" Mo Yan's writings during pretraining, for example, which could lead to lower perplexity for Mo Yan's tokens. Given that GPT tokenizes Chinese by individual characters, we sampled 500-character sequences from all four classes. The sampling process involved random selection of 3,000 sequences from each class to ensure the unbiased representation of texts.

The results (Figure 1) show that Maospeak features

[1] *https://spacy.io/models/zh*

[2] **Wenzhong 2.0**, 3.5B parameters, pre-trained on the 300GB version of the Wudao Corpus which includes mostly internet-based content (Zhang et al., 2022).

[3] **Chinese GPT2-xlarge**, 1.5B parameters, pre-trained on the 14GB CLUECorpusSmall corpus which includes news, Wikipedia, and social media content (Zhao et al., 2023).

[4] **CKIP GPT2 base-chinese**, 102M parameters, pre-trained on traditional Chinese data including Wikipedia and the CNA (Central News Agency) corpus. We have used the Python package `OpenCC` for simplified-traditional conversion.

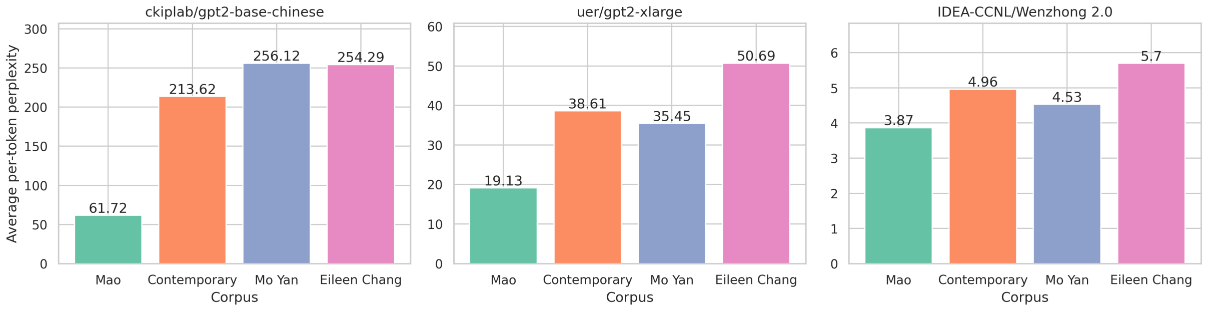


Figure 1: Average per-token perplexity across different GPT models.

a much lower perplexity than the other three classes, indicating a higher level of conformity with the language models' training data. By contrast, Eileen Chang's writings exhibit a very high per-token perplexity in all three models, reflecting more unexpected and creative word choices. The relatively high perplexity, and thus unpredictability, of Mo Yan's writing raises a question to consider for his critics, whereas the low perplexity of Maospeak hints at an important aspect of engineered languages (Ji, 2003), which promote the use of stock phrases and discourage creative usage of words (coining new metaphors, using rare vocabulary, unconventional syntax, etc). From this perspective, Maospeak resembles "machine text" rather than "human text," a distinction elaborated by Holtzman et al. (2019) in their work on neural text degeneration.

2.3 Systematicity

Another stylometric feature that differentiates texts from different authors is systematicity. Our hypothesis is that an author characterized by a high degree of systematicity would manifest a consistent overarching idea across all of their works. Essentially, each piece of a highly systematic writing can be viewed as a "microcosm" reflecting the broader semantic "macrocosm," even though individual texts may employ varied vocabulary.

One possible approach to measuring systematicity is to compute the average divergence between the overall ("global") vocabulary distribution across all texts produced by a given author and the ("local") vocabulary distribution in each of their specific writings. This methodology shares similarities with authorship attribution techniques such as z-scores of function words understood as an author-specific signal (Evert et al., 2017) which can be compared with a text-specific distribution through distance measures. However, our goal here is not to identify the real author among many possible ones, but to measure the particular author's thematic coherence.

The Kullback–Leibler (KL) divergence between the two probability distributions P and Q is given by the formula:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{i} P(i) \cdot \log\left(\frac{P(i)}{Q(i)}\right) \qquad (2)$$

For discrete probability distributions P and Q, the KL divergence quantifies the amount of information lost when Q is used to approximate P. In other words, if an event is likely under P but unlikely under Q, the term $\log\left(\frac{P(i)}{Q(i)}\right)$ is large, contributing significantly to the overall KL divergence.

To calculate the KL divergence for our datasets, we randomly sampled 3,000 segments from each of the four classes, with each segment containing 500 words. Here, P represents the local distribution of words in a specific segment, while Q is the global distribution of words across the entire corpus corresponding to the given class, including the unsampled fragments. For each class, the global and local vocabularies are uniquely determined within that class, i.e., there is no one shared vocabulary built at the outset.

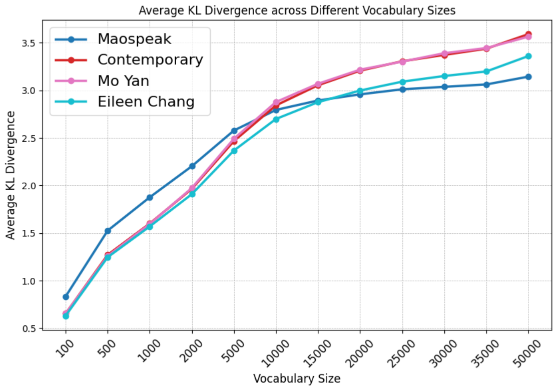


Figure 2: Average KL divergence across different vocabulary sizes; scale modified.

A few preliminary results encouraged us to conduct a series of tests and observe how the KL divergence changes as a function of vocabulary size. Surprisingly, Maospeak features the highest divergence for relatively small vocabulary sizes, which clearly separates it from the other three classes. This superiority diminishes, however, as we increase the vocabulary size (Figure 2). We confirmed the same results for segments of other lengths (100 and 1,000 words). A possible explanation of this behavior is as follows: when the average KL divergence continues to grow with the increasing vocabulary for the other three corpora (Mo Yan, Eileen

Chang, and the Contemporary corpus), it suggests that those texts feature a wide and diverse range of topics and themes. Each increase in vocabulary size continues to uncover more disparity between local and global distributions, which could potentially signify that these corpora are rich in specialized terminology or have a diverse set of topics or themes covered within them, and that less-frequent terms are distributed less uniformly across the 500-word segments.

By contrast, when the growth of divergence slows down with increasing vocabulary size, it may imply that the given corpus is more homogeneous and that most of the diversity or variability in word usage is captured at a smaller vocabulary size. Beyond a certain point, increasing the corpus-specific vocabulary size does not contribute significantly to revealing new disparities between local and global distributions. This suggests that the content of the Maospeak corpus is more focused and limited to a few main themes or topics, less-frequent terms being distributed more uniformly. At higher vocabulary sizes, Mao-style prose embodies the famous adage that "one sentence equals thousands of sentences" (一句顶一万句). The more we read Mao, the less we need to read Mao, since the amplitude of divergence is relatively small, whereas in the realm of literature every novel creates a new unexpected world. This last point holds as true for Mo Yan as for any other contemporary Chinese writer.

The above results suggest that it is not enough to compare the global-local divergence at a single vocabulary size. Eileen Chang, for example, exhibits the lowest divergence for middle-range vocabulary among all four classes, suggesting the relative coherence and internal similarity of her works, but her sentences and phrases become less alike once we take a larger vocabulary into account. In this sense, systematicity is an author-specific function of vocabulary size.

2.4 Characteristic words

Yet another method of measuring explainable differences in literary styles is to classify texts based on the presence or absence of characteristic words and expressions. TF-IDF, short for Term Frequency-Inverse Document Frequency, is a numerical statistic that reflects how important a word is to a document in a corpus. It is defined as follows:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \cdot \text{IDF}(t, D) \qquad (3)$$

Where:

- t is a term (or word)
- d is a document containing the term
- D is the corpus or collection of documents

The Term Frequency (TF) is calculated as:

$$\text{TF}(t, d) = \frac{\text{Count of term } t \text{ in } d}{\text{Total terms in } d} \qquad (4)$$

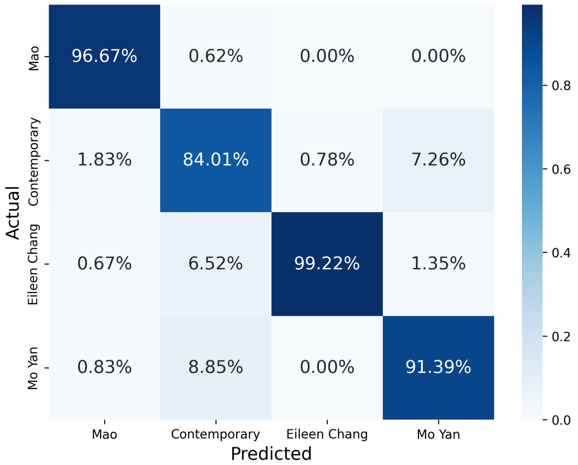


Figure 3: Confusion matrix for the Random Forest classifier with 300 TF-IDF features. Values have been normalized over the predicted conditions (columns).

The Inverse Document Frequency (IDF) is calculated as:

$$\text{IDF}(t, D) = \log\left(\frac{\text{Total docs in } D}{\text{Docs with term } t}\right) \qquad (5)$$

In other words, if a term t is frequent locally (4) and rare globally (5), its TF-IDF in the given document will be large. By utilizing TF-IDF, we emphasize words that are unique to a particular document while giving less weight to words that are common throughout the entire corpus. Training an explainable classifier, like a Decision Tree or Logistic Regression model, on such terms, enables the identification of features that most strongly indicate a particular style. In contrast to the two previous experiments, the features discovered in this way are relational and thus do not tell us anything about the particular literary style "as such," providing instead a way to distinguish different styles from each other.

The dataset in this experiment was built by dividing the four corpora into 500-word segments and then sampling 3,000 segments from each of them without replacement. In cases where fewer segments were available, we did not oversample. For example, given the smaller size of Eileen Chang's corpus, only 1,502 segments were obtained. All of the sampled fragments were put together and split into training (80%) and test sets (20%). We then trained a Random Forest classifier with 100 trees (estimators), each with a maximum depth of 15, the 300 words with the highest TF-IDF values obtained from the training data serving as our feature set. TF-IDF values were computed using the `TfidfVectorizer` from `scikit-learn`. We achieved 91.5% accuracy on the test set (2,092 examples), suggesting a relatively high degree of reliability in distinguishing different forms of writing, in particular those of Mao Zedong and Eileen Chang (Figure 3).

To gain deeper insights into which words are most indicative of a particular literary style, SHAP (SHapley Additive exPlanations) values have been computed post-training. SHAP values allow for the measurement of the impact (the average marginal contribution) that each feature (in this case, a word) has on the model's output (Mosca et al., 2022). For example, the computed SHAP values can reveal which words have the most influence in classifying a text as coming from the Mao Zedong corpus, essentially pointing out the vocabulary that distinguishes Maospeak from other styles.

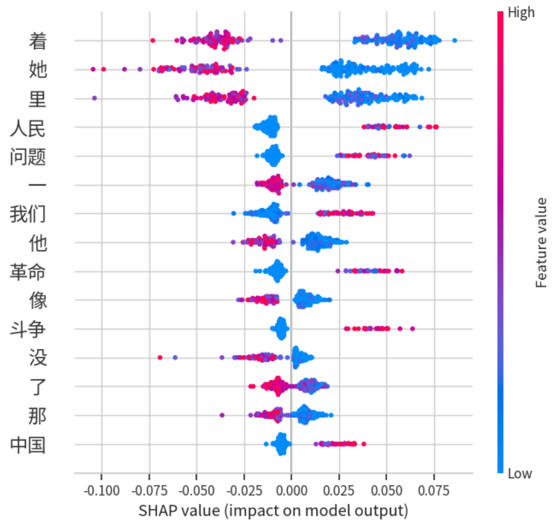


Figure 4: SHAP Values for class "Mao" in the Random Forest classifier, 200 test samples.

As shown in Figure 4, the computed SHAP values locate the main difference between Maospeak and the world of literature in the point-of-view markers: whereas literary texts are characterized by pronouns ("he" 他, "she" 她, "you" 你) and grammatical particles ("-ing" 着, "-ed" 了), which ground narratives in the actions and thoughts of individual characters, Mao-style prose "speaks" on behalf of the first-person plural "we" (我们) and gathers depersonalized, political terms such as "the People" (人民), "China" (中国), or "struggle" (斗争). Crucially, what the SHAP values demonstrate is that literary style is not only defined by the features present in a text but also by those that are absent, as shown by the blue dots which contributed (when absent, i.e., when bringing the TF value to zero or close to zero) to the model's final predictions. In this sense, it is the lack of point-of-view markers that characterizes Maospeak and the lack of political terms that characterizes contemporary prose. The role of absence within authorial signal is often overlooked by stylometric interpretations focusing solely on what is visible in the text, rather than what is not.

3 Conclusion

In this paper, we have analyzed three different aspects of Mao-style prose: perplexity, systematicity, and words with the highest TF-IDF values. The results of these experiments demonstrate some of the important features of Maospeak, an engineered language which reinforces the ideological tenets of Maoism through its formal characteristics. The conducted experiments also offer partial evidence to question the alleged Maoist influences on Mo Yan. While his violent literary style reflects China's revolutionary experience, it is hardly comparable to the redundant Party parlance.

While our analysis pertains chiefly to Maoism, we believe that our findings will be applicable also to more recent contexts, in China and beyond (Barmé, 2012). In particular, evaluations of next-token perplexity and KL divergence underscore the pivotal role of originality and subjectivity in language use. From this perspective, fiction reading and humanistic education become especially important. Reading widely and increasing one's exposure to various language data counters the influences of ideologies on our linguistically mediated perceptions of the world and increases the perplexity of our imaginations.

Limitations

Measurements of next-token perplexity are constrained by the availability of advanced hardware, the accessibility of large language models, the types of these models, and the amount and types of data that these models have been pre-trained on. In particular, tokenization proved a crucial consideration behind our choice of GPT models. In contrast to character-level tokenization used by Chinese versions of GPT, other tokenizers such as SentencePiece, used by generative models CPM (Zhang et al., 2020) and Chinese LLaMA (Cui et al., 2023), treats certain multi-character words (社会 "society" or 资本 "capital," e.g.) as single tokens, some of which are more prevalent in Mao's corpus compared to other corpora like that of Eileen Chang. In our tests, such discrepancies impacted measurements of perplexity, as the multi-character words are on average much less likely (i.e., they score lower probabilities) and thus increase the overall perplexity. Although the character-level tokenization of GPT models avoids this bias, treating each Chinese character individually and thereby providing a more uniform analysis across different writing styles and corpora, our choice of the pre-trained GPT models had a direct impact on the final results. Further analysis is needed to compare different models, tokenizers, and training data.

Acknowledgments

The author would like to thank the anonymous reviewers for their comments and suggestions. Special thanks are due to Aaron Gilkison and Heidi Huang for their feedback and support.

References

Geremie R. Barmé. 2012. New China Newspeak 新华文体. *China Heritage*. Online; accessed

09/25/2023.

J. F. Burrows. 1987. Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style. *Literary and Linguistic Computing*, 2(2):61–70.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca.

Stefan Evert, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. 2017. Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities*, 32:4–16.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *CoRR*, abs/1904.09751.

Libo Huang and Xinyu Shi. 2022. A corpus-based investigation of the english translations of mao zedong' s speeches. *Frontiers in Psychology*, 13.

Fengyuan Ji. 2003. *Linguistic Engineering: Language and Politics in Mao's China*. University of Hawaii Press.

Rafal Józefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *CoRR*, abs/1602.02410.

Adam Kilgarriff and Tony Rose. 1998. Measures for corpus similarity and homogeneity. In *Proceedings of the Third Conference on Empirical Methods for Natural Language Processing*, pages 46–52, Palacio de Exposiciones y Congresos, Granada, Spain. Association for Computational Linguistics.

Charles Laughlin. 2012. What Mo Yan's Detractors Get Wrong. *ChinaFile*. Online; accessed 09/25/2023.

Tuo Li. 1998. 汪曾祺与现代汉语写作——兼谈毛文体 [Wang Zengqi and Modern Chinese Writing: Also on the Style of Mao's Writings]. *花城 [Huacheng]*.

Perry Link. 2012. Politics and the Chinese Language: What Mo Yan's Defenders Get Wrong. *Asia Society*. Online; accessed 09/25/2023.

Perry Link. 2013. *An Anatomy of Chinese: Rhythm, Metaphor, Politics*. Harvard University Press.

Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. 2022. SHAP-based explanation methods: A review for NLP interpretability. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4593–4603, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Frederick Mosteller and David L. Wallace. 1963. Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302):275–309.

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying stylometry techniques and applications. *ACM Comput. Surv.*, 50(6).

Weihan Ou, Steven H. H. Ding, Yuan Tian, and Leo Song. 2023. Scs-gan: Learning functionality-agnostic stylometric representations for source code authorship verification. *IEEE Transactions on Software Engineering*, 49:1426–1442.

Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. 2016. Character-level and multi-channel convolutional neural networks for large-scale authorship attribution. *CoRR*, abs/1609.06686.

Yunita Sari, Mark Stevenson, and Andreas Vlachos. 2018. Topic or style? exploring the most useful features for authorship attribution. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 343–353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Michael Schoenhals. 2007. Demonising Discourse in Mao Zedong's China: People vs NonPeople. *Totalitarian Movements and Political Religions*, 8(3):465–482.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2014. Authorship attribution with topic models. *Computational Linguistics*, 40(2):269–310.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Anna Sun. 2012. The Diseased Language of Mo Yan. *The Kenyon Review*. Online; accessed 09/25/2023.

F. J. Tweedie, S. Singh, and D. I. Holmes. 1996. Neural network applications in stylometry: The "federalist papers". *Computers and the Humanities*, 30(1):1–10.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

Zhengyan Zhang, Xu Han, Hao Zhou, Pei Ke, Yuxian Gu, Deming Ye, Yujia Qin, Yusheng Su, Haozhe Ji, Jian Guan, Fanchao Qi, Xiaozhi Wang, Yanan Zheng, Guoyang Zeng, Huanqi Cao, Shengqi Chen, Daixuan Li, Zhenbo Sun, Zhiyuan Liu, Minlie Huang, Wentao Han, Jie Tang, Juanzi Li, Xiaoyan Zhu, and Maosong Sun. 2020. Cpm: A large-scale generative chinese pre-trained language model.

Zhe Zhao, Yudong Li, Cheng Hou, Jing Zhao, et al. 2023. Tencentpretrain: A scalable and flexible toolkit for pre-training models of different modalities. *ACL 2023*.

Efficient and reliable utilization of automated data collection applied to news on climate change

Erkki Mervaala and **Jari Lyytimäki**

Finnish Environment Institute Syke, Finland
firstname.lastname@syke.fi

Abstract

Automated data collection provides tempting opportunities for social sciences and humanities studies. Abundant data accumulating in various digital archives allows more comprehensive, timely and cost-efficient ways of harvesting and processing information. While easing or even removing some of the key problems, such as laborious and time-consuming data collection and potential errors and biases related to subjective coding of materials and distortions caused by focus on small samples, automated methods also bring in new risks such as poor understanding of contexts of the data or non-recognition of underlying systematic errors or missing information. Results from testing different methods to collect data describing newspaper coverage of climate change in Finland emphasize that fully relying on automatable tools such as media scrapers has its limitations and can provide comprehensive but incomplete document acquisition for research. Many of these limitations can, however, be addressed and not all of them rely on manual control.

1 Introduction

Despite the digital era's advancements, manual data collection continues to dominate humanities and social science studies, notably in media studies where the significance of digital communication is ever-increasing (Shearer and Mitchell 2021). Most print newspapers publish also online versions of their news content, and these online versions have exhibited modest variations in content compared to their print counterparts (Hoffman 2006; Mensing and Greer 2013; Hagar and Diakopoulos 2019).

The growth of online data has spurred the development of various automated data collection tools, such as media scrapers and public APIs, enhancing accessibility to vast datasets (Sirisuriya 2015; Aitamurto and Lewis 2013). However, the ease of collecting big data has potentially overshadowed inherent biases and errors, leading to availability bias and other types of bias affecting dataset representativeness (Mahrt and Scharkow 2013; Grimmer et al. 2022).

While web scraping is often viewed as a technical phenomenon, there is a growing discourse on the "softer issues" surrounding it, including ethical and legal considerations (Murray State University et al. 2020; Khder 2021; Zimmer 2010; Bruns 2019). The field is evolving, especially as platforms like Meta and Twitter have restricted data access.

Research on automated data collection has proliferated since the turn of the millennium, focusing largely on social media content (Scharkow 2013; Venturini and Rogers 2019). However, less attention has been given to utilizing automated methods for newspapers, with warnings about the trade-offs between automation and reliability (Deacon 2007; Mahrt and Scharkow 2013; Wijfjes 2017).

Media content analysis has traditionally involved small samples and qualitative approaches due to labor-intensive collection and coding. The shift towards automated research methods is motivated by the potential for larger sample sizes, despite reliability trade-offs (Broersma and Harbers 2018; De Grove et al. 2020; Wijfjes 2017; Blatchford 2020). Challenges and caveats related to computational methods, including supervised machine learning, have been discussed, emphasizing the need for caution in overestimating the benefits of automation (De Grove et al. 2020).

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 82–91
December 1-3, 2023 ©2023 Association for Computational Linguistics

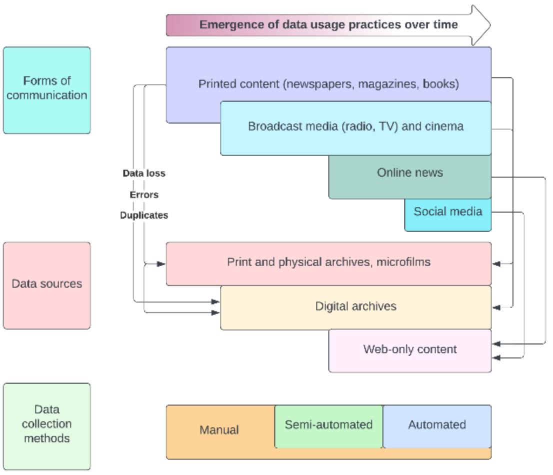

Emergence of data usage practices over time
Forms of communication
Printed content (newspapers, magazines, books)
Broadcast media (radio, TV) and cinema
Online news
Social media
Data loss
Errors
Duplicates
Data sources
Print and physical archives, microfilms
Digital archives
Web-only content
Data collection methods
Manual
Semi-automated
Automated

Figure 1: Evolution of data usage for media studies. The figure expresses data sources and usages of different media.

Media studies often lean towards manual or semi-automated collection methods, with less emphasis on fully-automated tools or "theory-driven online scraping" (Lodhia 2010; Khder 2021).

In Figure 1, we summarize the evolution of data usage and data collection methods and issues related to the reliability of data archiving from platform to platform. The aim of this article is to critically examine the pros and cons of different data collection methods and the crossing from manual and semi-automated data collection to fully automated practices. It is based on a case study focusing on newspaper data on climate change, showing the development of climate change news from 1990 up to December 2020.

2 Methods and materials

Our focus is on the news coverage of climate change in the Finnish newspaper Helsingin Sanomat (HS), given its high societal relevance, interdisciplinary character, and extensive previous studies on its climate coverage (Suhonen 1994, Lyytimäki 2011, Kumpu 2016, Teräväinen et al. 2011, Ylä-Anttila et al. 2018, Boykoff et al. 2019, Lyytimäki 2020). HS, the most widely circulated newspaper in the Nordic countries, has been a key source for monitoring media coverage of climate change in 58 countries and is a common subject in digital humanities and media studies (Boykoff et al. 2022).

The manual data (MD) for comparison comprises 14,750 news stories headlines retrieved from HS's online archive, spanning from January 1st, 1990, to December 31st, 2020. These stories, collected into a spreadsheet, were identified using specific climate-related queries (search screening full texts and using Finnish search terms for climate change, warming of climate and greenhouse effect) and included even those items mentioning climate issues tangentially (Lyytimäki 2011, 2015, Lyytimäki et al. 2020). Duplicates and irrelevant hits were removed based on manual inspection. Various factors, such as changes in the newspaper structure and search engine properties, influenced the data's format and content, with different information available across years and some data, like cartoons and advertisements, excluded.

Automated data were obtained using two different scrapers utilizing the Sanoma API. The first scraper (S1) mimicked the manual approach, collecting data in batches of 50 articles, mimicking the batch size of articles the manual online search provides after each click of the "show more" button, from oldest to newest, including full texts where possible, using the newspaper3k Python package. The second scraper (S2), based on the Finnish Media Scrapers project (Mäkelä and Toivanen 2021), performed 93 queries to the API, breaking down the search period into weekly segments and yearly intervals for each query term. As the manual dataset consisted only of headlines, publication dates and the article urls, the scrapers were set to collect only those data.

Both scraped datasets underwent cleaning to remove exact duplicates and ensure uniform formatting. The final comparison between manual and scraped datasets involved further cleaning and unifying data formats, focusing on the months the articles were published.

It is crucial to recognize that while MD, S1, and S2 all access the same news archive, the methodologies employed by each distinctly shape the dataset's composition. This underlines the significance of the data collection process itself, as it inherently filters and frames the information extracted from the archive. Therefore, any disparities in the collected data are attributed to the differences in collection methods and the inherent biases each method may introduce, rather than variations in the source material except in the cases when changes had been made to the archive's content or categorization in the times between the manual and scraped data collection.

While we acknowledge that inherent differences in the approaches of MD, S1, and S2 methods may lead to variations in the collected data, the comparison aims to highlight the nuances and potential biases each method introduces. The objective is to understand the trade-offs between manual and automated data collection, aiming to highlight the nuanced insights each approach offers and the unique biases they may introduce to the research on newspaper articles.

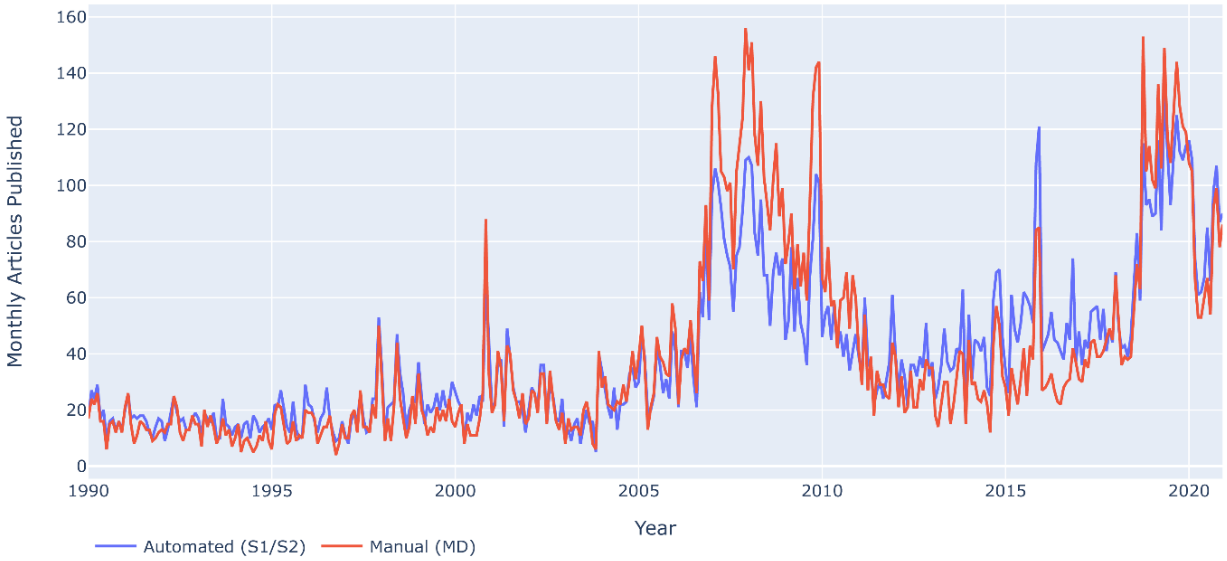


Figure 2: Articles on climate change published on Helsingin Sanomat 2000 – 2021 collected from online archive. The figure shows clear peaks in the frequency of climate change coverage but also highlights differences between the datasets.

3 Results

Compared to the manual dataset (MD) of 14750 news articles, neither of the datasets collected via the automated scrapers gave the exact same result. Also, different scraping techniques resulted in different amounts of articles. ¨

The S1 scraper queries resulted in 8227 stories on climate change, 7441 stories on greenhouse and 1576 on climate warming. After removing duplicates there were 14669 news articles published between January 3rd 1990 and December 31st 2020. The first article of the dataset details record heat in England and the last headline of the dataset declares that the year 2020 was the warmest year on record in Finland.

Initially, the S2 scraper provided the least amount of results: 7970 stories on climate change, 7437 on greenhouse and 1575 on climate warming with a total of 14553 articles after removing duplicates. Representing both retrieval and resource bias (Grimmer et al., 2022), the reason for the scraper collecting fewer articles than the other two is that the scraper ran into problems with either broken articles, manifesting as blank pages or error messages, or articles consisting of dynamic content that prevented scraping the full texts of the articles. After correcting this and limiting the results to article headlines only, S2 resulted in an almost identical result as the first scraper with only one article more, on climate change, than S1. From here on, we will discuss only the S1 dataset.

The full manual dataset of 14750 articles had 81 articles more than the 14669 of S1 (See Fig 1). While the difference between the datasets is only half a per cent in total numbers, the differences become more apparent when comparing certain peaks in the data: In November 2000 S1 dataset showed 69 published articles and MD 88 articles. Other similar peaks include February 2007 (S1: 106, MD: 146) and February 2008 (S1: 109, MD: 156). From 2011 to 2018, the S1 seems to take over and contain more results. The largest peaks of S1 align with the December 2015 Paris Accord when S1 displayed 121 results and MD only 85. From 2018 to the beginning of 2020, MD displays more results on average and after that S1 again until the end of the year 2020.

On closer inspection, including a detailed manual review of the discrepancies, focusing on the type and content of articles that differ between the datasets, the articles causing the differences are mainly smaller commentaries, opinion pieces or editorials, and on a smaller scale, television or radio programming details. For December 2007 MD has 156 articles and S1 had 109 articles. The differences appear to come from more opinion piece articles included in the manual dataset compared to the scraped set. While some opinion pieces and editorials were included in the scraped set, MD included numerous relevant ones such as a small comment piece titled “Vuoden viherpesu” (“Green Wash Of The Year”).

In the opposite case of December 2015, the surplus of articles in the scraped dataset is mainly the result of several different editions of the same story published on two different sections of the site such as “ulkomaat” (“foreign”) and “ilta” (“evening”). In addition, some opinion pieces were included in the scraped set that were not present in the manual set.

When calculating the percentage of matching articles between the datasets, using their unique identifiers, the article headlines and urls, the datasets were only 84,2 % identical. The differences can be mostly explained by differences in coding the articles in the manual set and the automatically retrieved headlines from the online archive which in turn may also change over time especially if the articles were subjected to A/B testing, usually changing the articles’ headlines to optimize online readership, during or after the data collecting. It should be pointed out that in February 2023 an editor of Helsingin Sanomat admitted to modifying headlines of their online and print versions differently and an editor of the evening tabloid Iltalehti stated that negative headlines work better as they interest people more (Sillanmäki, 2023).

These kinds of discrepancies should, however, be also accounted for when assessing different ways of obtaining data. A more reliable way to compare articles would be to use the articles’ hyperlinks that are not likely to change over time.

Considering a stricter approach to removing duplicates, some articles were indeed almost identical to each other when it comes to the headline and even the article content despite having different hyperlinks. Removing duplicates based solely on the title or solely on the hyperlink may still leave different versions of the article in the datasets as some archived articles from the beginning of the datasets’ time period may have both the print version and the online version of the article available online with individual hyperlinks with minor variations in the online headline. In some cases, the same article was published twice within the same month with a different hyperlink. Also, the same or very similar headlines may lead to a “full” and an “abridged” version of the story. A combination of filtering by unique hyperlinks and headlines with the possible addition of publication month and content comparison may be a more accurate, though more cumbersome, approach.

4 Discussion

4.1 Automation as a solution

Updates in search engines and content and categorizations of the database may distort search results updating old data. It is also possible that some items related to climate issues are missing from the sample because of the limited set of keywords. Therefore, it is vital to conduct test searches to ensure that the right balance is found between exclusion and inclusion. This, in turn, requires expertise on the qualities of the issues under scrutiny. For example, coverage of biodiversity loss and “the polycrisis” may overlap with climate change coverage.

While manual data collection can offer a relevancy filter of sorts already during the collecting process, it is slow as all the details of the articles have to be manually copied and pasted or written in the data set document. The manual collecting process raises also issues with repeatability and handling errors in the original tasks found out later during the process. Especially with vast datasets, noticing an error after the data has been collected, it may not be possible to repeat the process afterwards due to limited human resources. The speed of automated data collection depends mainly on the processing power attributed to the scraper and the amounts of articles published during the period in question. For example, scraping article headlines for the search query “climate change” can take anything between a few seconds to a few minutes. For manual collection, the time spent can be considerably longer (Lauer et al. 2018), often beyond the resources available. Although automated scraping significantly enhances cost-efficiency and data breadth, it is not without trade-offs. For instance, automated methods may inadvertently capture irrelevant data, necessitating post-collection filtering that can be both labor-intensive and prone to oversight. This underscores the importance of a balanced approach that weighs the speed and scope of automation against the precision and context sensitivity of manual data collection..

The automated method also offers the possibility to collect much larger datasets much quicker and therefore the possibility of more comprehensive scopes for studies even if the data would have to be filtered down later. Manual collection can also suffer from a lack of timeliness as collecting the data can be too slow to produce data fast enough for topical analysis on quickly evolving topics. Apart from the comparable slowness, additional human errors and biases can be coped with via well-established ways such as intercoder reliability tests.

Relying on automated methods may easily lead to omissions in reliability testing as data collected automatically can be assumed to have been collected "objectively". In order to find the most reliable solution, testing between different automated methods and comparing results to similarly produced manual samples would be one way to address this issue, albeit time-consuming. The need for such testing increases with the gaps between data collection sessions as changes in APIs may result in different search results.

. Especially with larger datasets consisting of thousands or millions of data points, systematic errors, that might have been caught more easily by human eyes, may go unnoticed by the researcher relying on automated data collection. Therefore, testing the methodology via smaller test runs is encouraged. While a scraper can perform perfectly fine for 90 per cent of the news articles, the remaining ten per cent may cause issues for the whole dataset. For example, a single misplaced comma or a semicolon scraped in the scraped data may mess up the following rows and columns. Additionally, especially on archived content, the scraper may hit a wall due to bad or obsolete programming. Such issues arise most often when scraping for full articles as each news story is a page of its own for the scraper to run into error-inducing content which at best may lead to empty content cells in the dataset. For these reasons, error handling is very important in the scraping process.

Causes for such systemic errors can also change over time. For example, changes in the newspaper website infrastructure such as adding CAPTCHA, a program that checks whether the user is human or a machine, and other anti-scraper measures will affect the results and possibly prevent for example collecting full texts of articles especially if the articles themselves are behind a paywall. Additionally, the introduction of the so-called "dynamic articles" that feature semi-interactive and interactive elements that reveal text as the reader scrolls down the article, also affects collecting the full texts of the articles, as they often require more sophisticated scraping techniques, frequently requiring site-specific programming. Such dynamic articles may be challenging for manual data collection as well.

Finally, there are possible issues with timestamping the data. As the data is for the exact times when the articles were published and modified are available via scraping, there is a need to normalize the ordering of the data in the dataset whether it be by year, month, day or by minute. Whereas in fast-paced social media communication it may be important to know the publishing time by the second, in online news media analysis the timestamping may not need to be as detailed. The article can also be modified or republished after its original publication which may lead to the article being misplaced in the dataset depending on which variable one uses to sort articles by – for example "time published" or "time modified". Though an issue of potentially limited relevance, should an article be updated for instance at the change of a month, it may be duplicated in a collection of datasets updated monthly. Additionally, the order of the articles may be relevant for consequential articles covering short-lived, fast-paced events.

4.2 Common challenges

There are also several common challenges for both manual and automated data collection. Changes in visual design and composition of the sections of the newspaper may have an influence on the number, length, and presentation style of news items. For example, during the study period, the composition of the printed version of HS was renewed several times, including a major change from broadsheet to tabloid on 8 January 2013. (Sanoma 2012). The data itself may not be complete as the provider may have altered the archive over the years. These kinds of archive alterations may not have had any nefarious intentions behind them as they may have been part of restructuring the archive for better accessibility or functionality and may be limited to actions such as removing duplicates or recategorizing content. In some cases in the HS dataset, duplicate versions of articles were found even with a different hyperlink as they represented different versions such as online and print versions of the same article with only minimal changes.

Proper (automated) comparison of the manual and scraped datasets require some unification and cleaning for the data. As the manual collecting process for large datasets often includes more than one researcher and may stretch to long periods of time, differences in recording the data are bound to be more frequent compared to automated scrapers that perform the task without variations. Omitted details can for example be added to the manual datasets using even the same automated tools used for scraping. It should be noted that each comparison case is different, and the methods and tools required to address such issues should be assessed by case and by data type.

The transformation of news media from static text to dynamic, multimedia narratives presents both opportunities and challenges for data collection. Visual elements like photographs, infographics, and videos are integral to modern storytelling and can significantly influence audience perception. However, these non-textual elements are often not captured by traditional scraping techniques, highlighting a gap in our methodology that future studies will need to bridge to fully understand media impact.. Additionally, in recent years we have seen an uptick in different kinds of more complex news content such as the aforementioned dynamic news articles, and interactive news articles with sliders, polls and calculators, both providing valuable journalistic content and even significant amounts of text data to the reader but more complex to include as part of a text-based study. Embedded content may also prove to be difficult to access in the future, especially if it is included content that has since been deleted from the source. Deleted Tweets from Twitter/X, for example, are not accessible via those news articles that have embedded them in the middle of the news text after the deletion. Even though the contents of such Tweets would have been written out within the news text, they often are not verbatim and, if not in the native language of the publication, are translated.

These issues reflect the overall evolution of a news article and the structural changes of news over time. Are both a long-form written piece and a news item including infographics and info boxes considered individual news stories? What about stories that are ever-changing or constantly updated such as articles following the global carbon budget diminishing every minute or articles related to the COVID-19 pandemic with daily updates on infections and victims? One way to individualize an article could be based on the article's hyperlink. Then, if the article is changed, the hyperlink stays the same. This, however, does not take into account the potential changes in the message the article conveys to the reader. An article's headline can change several times during the day of the publication due to click optimization, A-B testing, and localization to name a few reasons (Hagar and Diakopoulos 2019). The "original" headline could be said to be the one appearing on the paper version of the newspaper but then articles without a printed counterpart would have to be omitted.

It is therefore paramount for the transparency and reproducibility of the data that a timestamp of the data collection is included also in the dataset. As changes and corrections in the text are often highlighted in the articles in question after the fact, the timestamp, while not covering the change, can at least indicate whether the article was included in the dataset before or after the alteration.

The issue with the changing headlines is a recent one but an important one. While we do not focus on the messaging and framings in the headline in this article, the changes made to headlines that appear to the readers in different forms over different times, devices and platforms is an important topic for media studies and would have to involve tools closely monitoring such changes. A similar approach could and should be applied to the changes in the content of the articles. In fact, there are some instances that already collect and publish changes in headlines and content of news publications online.[i]

4.3 Editorial decisions and the evolution of the language used

The caveats for any use of automated online search functions of newspapers include the possibility that there may be articles omitted from the dataset that could be argued to be categorized as related to a topic such as "climate change" but for some reason have not been included. These omissions could, however, be argued to represent in a rather transparent way the views of the news outlets. If an article is not included in the search results, whether on purpose or not, the media outlets communicate to their readers that the article in question is in fact not relevant in that context. The lack of categorization of the "missing articles" may, of course, have other, "human" reasons, too. The time and resource constraints at the media organization may play a role, as well as potentially the expertise dealing with the categorization, especially if done manually, may lead to the omission of some articles appearing relevant to climate scientists but perhaps not to the media in question. The primary category attached to the article may also be a factor, as several crises such as food shortages may in fact have to do with climate change but are not categorized primarily as such.

The historical topic relevancy is also a factor, and search strategies should allow comparisons between different times and places. Climate change provides an example of a global issue with shared key terminology across different contexts, but languages differ in their emphasis as exemplified by the lack of use of the term "global warming" in Finnish debate. The language used to describe climate change has evolved considerably over the years, which is apparent in the data as we look at the yearly datasets by the scraper search queries: in 1990 there were 18 articles categorized as "climate change", 16 articles as "climate warming", and 295 articles on "greenhouse*", respectively, while in 2020 the respective figures were 1052, 82, and 288. Not only did the amount of the articles increase but also the shift to using the term "climate change" ("ilmastonmuutos") instead of "greenhouse effect" ("kasvihuoneilmiö") is apparent. By sheer quantity, the switch seems to have happened between 2006 and 2007, which coincides with the publication of the influential Stern Review on the Economics of Climate Change (Stern 2007) released in October 2006.

Additionally, even if the news story on climate change has been categorized by a news outlet in the category "climate change", the article may still be omitted from search results with the search query "climate change" for some other reason unknown to the public. For example, recent climate coverage in Finland often deals with carbon sinks of the Finnish forestry not necessarily mentioning the term climate change and labelled under energy policy rather than climate policy. The same retrieval bias applies to the concept of "emissions" as relevant stories may include references to emission targets but not climate change specifically. Furthermore, the apparent easiness of using such digital databases may tempt simplification in framing a complex topic such as climate change and prompt conclusions omitting the context. Similar simplification has been found for example in the coverage of Africa (Madrid-Morales 2020).

All in all, the Finnish newspaper archiving system does offer a wide array of opportunities for research: Historical newspapers are comprehensively digitalized with public and free access as their copyrights have already expired. While there are no comprehensive digital archives for more recent media coverage, the consolidation of media companies has led to archives combining materials from some previously independent newspapers. In these cases, the availability of copyrighted materials depends on the right owner. Access to such easy-to-use digital archives may also limit the usage of a certain database over another. HS not only provides the digital archive from 1990 onwards but also an archive of digital replicas of their newspapers from 1889 to 1997 in PDF format. Full texts are made available for subscribers. The PDF archive is, however, not as easy to analyze via automation and machine learning and would require for example tools related to computer vision.

Finally, compared to research on print editions or their virtual counterparts such as PDF copies, online news archives are unable to provide information on the visibility given to the article on the day of publication. Though the front page of the print edition and the main stories on the web page do frequently differ, online news archives only tell when the story was been published with possible additions of its categorization and type.

5 Conclusion

Our findings reveal the impracticality of an exhaustive data collection strategy, challenging the notion that completeness equates to comprehensiveness. Instead, our research underscores the need for strategic sampling, where the focus is on capturing a representative swath of articles that collectively provide insight into the evolution and nuances of issues such as climate change coverage. Whether collected via automated scrapers or manual methods, it is very likely that all the news articles published will not be included in the dataset. There is a risk of complete lack and omissions of data for poorly deposited early years and risks related to diversifying presentation formats for recent years. Significant caveats should be addressed remaining caveats always communicated effectively.

In order to avoid the research methodology becoming a black box, we advocate for meticulous documentation of data collection processes. This includes detailing the algorithms, API settings, and decision-making criteria employed during data scraping. Such transparency not only enhances the reproducibility of research but also allows for a critical evaluation of the methodologies used, promoting trust and verifiability in the findings. This is not limited to only including timestamps for the collecting periods but also the selected settings/features/attributes of the APIs and other relevant scraper features used. Typically, there is a routine expectation for transparency regarding the process of subjective data collection, especially in human-based methods. However, this level of scrutiny is often overlooked when it comes to automated methods.

On the other hand, this responsibility could be shifted or partially shared if the data are not collected by the authors themselves but are provided by an external entity such as a company specialized in media analysis and scraping or even the news outlet itself. In the latter case, one then has to trust the outlet that they provide all the news stories on the topic they deem relevant. Additionally, in both the former and latter cases, the data collection becomes a true black box as reproducing the data collection is not possible based on solely the research article.

While our study concentrates on the frequency of climate change articles, we acknowledge that this is a mere slice of the narrative. The visibility and prominence given to these articles — such as front-page placement or feature positions on websites — play a crucial role in shaping public discourse. Future research could enrich our understanding by incorporating these dimensions, potentially utilizing sophisticated tools to analyze digital replicas and virtual formats for a more holistic picture of media influence.

Finally, we highlight the importance of securing public non-commercial databases collecting and storing media data. As media conglomerates and social media companies apply stricter commercially based data policies, such public databases become increasingly important both for manual and automated approaches.

References

Tanja Aitamurto and Seth C. Lewis. 2013. Open innovation in digital journalism: Examining the impact of Open APIs at four news organizations. New Media & Society, 15(2):314–331.

Ralf Barkemeyer, Frank Figge, Andreas Hoepner, Diane Holt, Johannes Marcelus Kraak, and Pei-Shan Yu. 2017. Media coverage of climate change: An international comparison. Environment and Planning C: Politics and Space, 35(6):1029–1054.

Annie Blatchford. 2020. Searching for online news content: the challenges and decisions. Communication Research and Practice, 6(2):143–156.

Max Boykoff, Meaghan Daly, Rogelio Fernandez Reyes, Jari Lyytimäki, Lucy McAllister, Marisa McNatt, Erkki Mervaala, Ami Nacu-Schmidt, David Oonk, and Olivia Pearman. 2019. World Newspaper Coverage of Climate Change or Global Warming, 2004-2023. Media and Climate Change Observatory Data Sets. Cooperative Institute for Research in Environmental Sciences, University of Colorado.

Maxwell T. Boykoff. 2011. Who Speaks for the Climate?: Making Sense of Media Reporting on Climate Change. Cambridge University Press, 1st ed.

Marcel Broersma and Frank Harbers. 2018. Exploring Machine Learning to Study the Long-Term Transformation of News: Digital newspaper archives, journalism history, and algorithmic transparency. Digital Journalism, 6(9):1150–1164.

Axel Bruns. 2019. After the 'APIcalypse': social media platforms and their fight against critical scholarly research. Information, Communication & Society, 22(11):1544–1566.

Frederik De Grove, Kristof Boghe, and Lieven De Marez. 2020. (What) Can Journalism Studies Learn from Supervised Machine Learning? Journalism Studies, 21(7):912–927.

David Deacon. 2007. Yesterday's Papers and Today's Technology: Digital Newspaper Archives and 'Push Button' Content Analysis. European Journal of Communication, 22(1):5–25.

Stacy Gilbert and Alexander Watkins. 2020. A comparison of news databases' coverage of digital-native news. Newspaper Research Journal, 41(3):317–332.

Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2022. Text as data: a new framework for machine learning and the social sciences. Princeton University Press, Princeton Oxford. ISBN: 9780691207544

Nick Hagar and Nicholas Diakopoulos. 2019. Optimizing Content with A/B Headline Testing: Changing Newsroom Practices. Media and Communication, 7(1):117–127.

Lindsay H. Hoffman. 2006. Is Internet Content Different after All? A Content Analysis of Mobilizing Information in Online and Print Newspapers. Journalism & Mass Communication Quarterly, 83(1):58–76.

Moaiad Khder. 2021. Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. International Journal of Advances in Soft Computing and its Applications, 13(3):145–168.

Ville Kumpu. 2016. On making a big deal. Consensus and disagreement in the newspaper coverage of UN climate summits. Critical Discourse Studies, 13(2):143–157.

Claire Lauer, Eva Brumberger, and Aaron Beveridge. 2018. Hand Collecting and Coding Versus Data-Driven Methods in Technical and Professional Communication Research. IEEE Transactions on Professional Communication, 61(4):389–408.

Sumit K. Lodhia. 2010. Research methods for analysing World Wide Web sustainability communication. Social and Environmental Accountability Journal, 30(1):26–36.

Jari Lyytimäki. 2011. Mainstreaming climate policy: the role of media coverage in Finland. Mitigation and Adaptation Strategies for Global Change, 16(6):649–661.

Jari Lyytimäki. 2020. Environmental journalism in the Nordic countries. In In David B. Sachsman, & JoAnn Myer Valenti (Eds.) Routledge handbook of environmental journalism, pages 221–233. Routledge, London and New York. ISBN: 9781032336442

Dani Madrid-Morales. 2020. Using Computational Text Analysis Tools to Study African Online News Content. African Journalism Studies, 41(4):68–82.

Merja Mahrt and Michael Scharkow. 2013. The Value of Big Data in Digital Media Research. Journal of Broadcasting & Electronic Media, 57(1):20–33.

Eetu Mäkelä and Pihla Toivanen. 2021. Finnish Media Scrapers. Journal of Open Source Software, 6(68):3504.

Donica Mensing and Jennifer D. Greer. 2013. Above the Fold: A Comparison of the Lead Stories in Print and Online Newspapers. In Internet Newspapers, pages 283–302. Routledge, 0 ed.

Murray State University, Vlad Krotov, Leigh Johnson, Murray State University, Leiser Silva, and University of Houston. 2020. Legality and Ethics of Web Scraping. Communications of the Association for Information Systems, 47:539–563.

Sanoma. 2012. Helsingin Sanomat to go to the tabloid format. Sanoma.com.

Andreas Schmidt, Ana Ivanova, and Mike S. Schäfer. 2013. Media attention for climate change around the world: A comparative analysis of newspaper coverage in 27 countries. Global Environmental Change, 23(5):1233–1248.

Elisa Shearer and Amy Mitchell. 2021. News Use Across Social Media Platforms in 2020. Pew Research Center.

Lotta Sillanmäki. 2023. HS:n päätoimittaja vastaa kritiikkiin verkon ja printin erilaisista otsikoista: ”Ei mennyt ihan putkeen”. Yle.fi.

S.C.M. de S Sirisuriya. 2015. Comparative Study on Web Scraping. Proceedings of 8th International Research Conference, KDU.

Nicholas Stern, editors. 2007. The economics of climate change: the Stern review. Cambridge University Press, Cambridge, UK ; New York.

Pertti Suhonen. 1994. Mediat, me ja ympäristö. Hanki ja jää, Helsinki. ISBN: 9518916446

Tuula Teräväinen, Markku Lehtonen, and Mari Martiskainen. 2011. Climate change, energy security, and risk—debating nuclear new build in Finland, France and the UK. Energy Policy, 39(6):3434–3442.

Tommaso Venturini and Richard Rogers. 2019. “API-Based Research” or How can Digital Sociology and Journalism Studies Learn from the Facebook and Cambridge Analytica Data Breach. Digital Journalism, 7(4):532–540.

Huub Wijfjes. 2017. Digital Humanities and Media History: A Challenge for Historical Newspaper Research. TMG Journal for Media History, 20(1):4.

Tuomas Ylä-Anttila, Juho Vesa, Veikko Eranti, Anna Kukkonen, Tomi Lehtimäki, Markku Lonkila, and Eeva Luhtakallio. 2018. Up with ecology, down with economy? The consolidation of the idea of climate change mitigation in the global public sphere. European Journal of Communication, 33(6):587–603.

Michael Zimmer. 2010. “But the data is already public”: on the ethics of research in Facebook. Ethics and Information Technology, 12(4):313–325.

[i] For example, there are several bot accounts on Twitter that highlight changes made to newspaper articles such as “Editing The Gray Lady” or @nyt_diff that reveals changes made on the main page of the New York Times website.

Unlocking Transitional Chinese: Word Segmentation in Modern Historical Texts

Baptiste Blouin[a], **Hen-hsen Huang**[b], **Christian Henriot**[a], **Cécile Armand**[a]
[a] IrAsia, Aix-Marseille University
first.last@univ-amu.fr
[n] Institute of Information Science, Academia Sinica
hhhuang@iis.sinica.edu.tw

Abstract

This research addresses Natural Language Processing (NLP) tokenization challenges for transitional Chinese, which lacks adequate digital resources. The project used a collection of articles from the *Shenbao*, a newspaper from this period, as their study base. They designed models tailored to transitional Chinese, with goals like historical information extraction, large-scale textual analysis, and creating new datasets for computational linguists. The team manually tokenized historical articles to understand the language's linguistic patterns, syntactic structures, and lexical variations. They developed a custom model tailored to their dataset after evaluating various word segmentation tools. They also studied the impact of using pre-trained language models on historical data. The results showed that using language models aligned with the source languages resulted in superior performance. They assert that transitional Chinese they are processing is more related to ancient Chinese than contemporary Chinese, necessitating the training of language models specifically on their data. The study's outcome is a model that achieves a performance of over 83% and an F-score that is 35% higher than using existing tokenization tools, signifying a substantial improvement. The availability of this new annotated dataset paves the way for refining the model's performance in processing this type of data.

Introduction

Previous studies of the Chinese language based on NLP methods have focused on either modern Chinese — today's Chinese for which there exists a wealth of digital resources such as Wikipedia, Baidu, etc. — or on classical or ancient Chinese based mostly on collections of literary, religious, or medical texts. In this paper, we address the issue of word segmentation for the Chinese language that emerged and developed between the end of the 19th century and the early 1950s. We shall label this language "transitional Chinese".

Despite the huge amount of publications that appeared in China during that century in the form of newspapers, periodicals, encyclopaedias, books, reports, etc., almost no work has been done to address the challenges that transitional Chinese raises for the implementation of NLP methodologies. The major reason is the absence of available digital resources. Although the major libraries and private companies in China have digitized a great number of newspapers or periodicals, the access to the digital versions, when they exist, has been limited to a web interface, most of the time with severe restrictions on downloading the text files or even on copying sections of the text. These databases are designed for text display, consultation, and reading, not for text analysis through computational methods.

Chinese companies and institutions as a rule refrain from providing the text files that would allow researchers to fully implement NLP methodologies. Despite repeated attempts to negotiate TDM rights (Text and data mining) with several mainland companies, we hit a wall when it came to obtaining the text files, even under the strictest terms of confidentiality and corpus protection. To put it bluntly, there is an abyssal gap between the extraordinary effort made at digitizing historical sources in China, especially the vast reservoir of periodicals and newspapers at the Beijing National Library or the Shanghai Library, and the possibility to use these resources to advance research.

The ENP-China project was designed from its inception with NLP methodologies as a key component of its methodology to process historical sources. The press, in particular, was considered as crucial not just because it could provide rich historical information, but also because this was the very

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 92–101
December 1-3, 2023 ©2023 Association for Computational Linguistics

place where the modern Chinese language developed. We were able to acquire the entire collection of the *Shenbao* in full-text files thanks to a private provider based in Taiwan. With this treasure trove, we were in a unique position to finally address the challenge of implementing NLP methodologies and algorithms trained on modern texts on a vast corpus of pre-1949 transitional Chinese as found in the press.

Our work represents the first attempt to break through the limits of existing models for Chinese and to develop models adapted to transitional Chinese through campaigns of annotations to create training sets with verifiable data. A major challenge was to design a model that proved robust across the various genres of texts found in the *Shenbao* and across the whole period under study (1872-1949). The central objective of our experiments and developments was to adjust models and to tailor them to this evolving Chinese — from administrative Classical Chinese to modern common and literary Chinese — with multiple purposes:

- to extract as completely and as accurately as possible the relevant historical information on our research topics
- to enable textual, discourse analysis at a scale heretofore unreachable
- to deliver new datasets for corpus and computational linguists

Newspapers are a most relevant source for analyzing long-term patterns of linguistic and conceptual changes (Hengchen et al., 2021). The newspaper that we used as a core resource represents a huge collection of more than 2.2 million articles published between March 1872 and May 1949. The *Shenbao* was the first daily newspaper published in Chinese in Shanghai. Originally a local publication, it became at once a national newspaper read throughout the empire. It also set the matrix for the subsequent newspapers that appeared at the turn of the century. For almost thirty years, the *Shenbao* set the tone, the pace and the model of news-writing, thereby creating a language in itself, the same language found in later publications (Mittler, 2004).

When the Shenbao was established in 1872 in Shanghai, there was no previous history of mass print media in China. The Chinese state and its local agencies produced ”official gazettes” that printed and circulated official decrees, edicts and lists of appointed officials. These texts were written according to the conventions of administrative classical Chinese which, despite some evolution, remained basically the same for the few centuries before the advent of the modern press. The establishment of privately-run newspapers, however, raised language issues:

- the newspapers needed to reach a wider educated audience than just the officials and literati and make the language of news reporting more accessible.
- the newspapers reported on a much wider range of issues that touched on new topics and notions, especially when it came to international news.
- the newspapers developed *sui generis* from classical Chinese a new language through successive shifts, both in vocabulary, grammatical structure, use of punctuation, layout, and typography.

The classical Chinese language differs considerably from modern Chinese. It consists mostly in single-character words. There are of course exceptions, which include the name of institutions, titles, proper names, etc. Yet, in most classical Chinese texts a character by and large equates a word. This feature generates a phenomenon of polysemy of characters that only the context in which they appeared and the ingrained knowledge acquired by the literati in the major genres (Confucian classics, poetry, memorials, etc.) of literary writing could disambiguate.

Newspapers faced several challenges in adapting the classical language to the constraints of news reporting.

- First, they had to introduce a large range of new expressions to cover in enough precise terms all the concepts and objects that came to China through its interaction with the outside world. Japan, that had been exposed earlier on to concepts imported from the West, became a major supplier of character-based neologisms (Wang, 1998; Chen, 2014). It consisted mostly in two-character expressions that eventually became the norm in modern Chinese. Whereas classical Chinese was quite strictly monosyllabic, modern Chinese became mostly disyllabic.

- Second, the grammatical structure of classical Chinese changed seamlessly in various ways, including a change of some of the basic elements such as pronouns, demonstratives, verbs, etc., while punctuation was introduced incrementally and sometimes a bit haphazardly (Mullaney, 2017). This process of change did not follow any guidelines. The Chinese language reinvented itself under the collective movement and innovation of the literati. Several initiatives by the state sought to define rules for the creation of a modern language, which probably had a certain impact, but it was not until 1922 that a new national language was officially adopted, to be taught throughout the entire school system (Kaske, 2008). Nevertheless, previous writing habits persisted until 1949, which make the newspapers of the Republican era a kaleidoscope of Chinese writing styles and languages.

- Third, the early newspapers included in their pages various genres of texts written in very different styles, from extracts from official gazettes, to literary texts, including poetry, to translations from fiction or telegrams, to news reporting and advertisements. Although some sections lost in importance with time (official gazettes), newspapers generally formed a mosaic of writing styles. The *Shenbao* presented such a kaleidoscope of overlapping genres (Mittler, 2004).

- Fourth, the period from 1872 to 1949 is one of progressive but constant language change. Even in 1872, except for the excerpts from the *Beijing Gazette* (京報) written in administrative classical Chinese, the language used in the *Shenbao* for news reporting was already a different language from the start. Through a data-driven analysis of the content of the *Shenbao*, Magistry has identified six main periods based on clustering. It is consistent with previous studies that defined 1904, 1911 and 1937 as clear-cut shifts. Magistry's study, however, points to other shifts around 1890-1892 and another one in 1922 (Magistry, 2021).

Word segmentation constitutes a prerequisite for many tasks of textual analysis such as topic modeling, word frequencies, semantic network analysis, etc. Whereas the task of word segmentation for languages based on Latin characters has become almost trivial, it remains a challenge for Chinese.

Whether in modern or in classical Chinese, characters are aligned vertically or horizontally next to one another. In the case of classical Chinese, all the way to the 1920s one also faces the near or complete absence of punctuation (Galambos, 2021). Although modern punctation was introduced in the last decade of the 19th century, its usage remained very uneven and unstable until the early 1930s (Hamm, 2021). While punctuation does not separate tokens per se, the presence of punctuation introduces a significant element of sentence segmentation that helps for tokenization. Modern Chinese has received a lot of attention, with a constant flow of papers in the major conferences on various methods to produce accurate word segmentation. Although less rich, works have also focused on classical (and even ancient) Chinese (Chen and Tai, 2009; Huang and Wu, 2018; Han et al., 2018). While both strands of research provide models, pretrained resources, and conceptual frameworks, neither procure readily replicable and usable models for transitional Chinese. The main challenge in the ENP-China project therefore has been to design a robust model that can adapt to the various stages of language development in transitional Chinese. In the next sections, we present the context of the study by reviewing previous appraoches to Chinese word segmentation and available datasets for training models. We introduce our new dataset, the experiments we have made with existing models and the model that we have trained.

1 Chinese word segmentation

Chinese word segmentation (CWS) has been a subject of prolonged debate within the field of NLP. For specific tasks and datasets, adhering to the character level proves to be a more suitable approach, yielding higher-performance results and simplifying the process. Thus, in certain scenarios, it is deemed less essential to perform word segmentation. For instance, a study by (Li et al., 2019) strongly asserts that with the advent of neural methods in NLP, CWS is gradually becoming an irrelevant or even detrimental step in the NLP pipeline.

In the context of computational methods in the humanities, the task of addressing CWS continues to be an area of significant interest. In fact, this crucial task enables multiple analyses of Chinese text, empowering researchers and practitioners to extract valuable insights and knowledge.

To accomplish this, it becomes essential to have a tokenizer that is specifically tailored and adapted to the data being processed. The effectiveness of the tokenizer greatly impacts the accuracy and efficiency of subsequent NLP tasks on Chinese texts.

Over time, CWS has been extensively studied, leading to the development of sophisticated systems capable of achieving near-perfect results. Some of these systems boast an impressive F-score close to 100%, indicating a high level of precision and recall in identifying word boundaries within Chinese texts. These remarkable achievements in segmentation accuracy further enhance the overall performance of NLP applications when working with Chinese language data.

During the early stages of CWS, the predominant methods employed were lexicon- and rule-based approaches, such as forward maximal matching, reverse maximal matching, and least word cut. While these techniques were relatively straightforward to implement, they suffered from low accuracy in segmenting words effectively.

To address this challenge, from a machine learning perspective, CWS started to be approached as a sequence labeling task. This method involved predicting whether each input character should be separated from its neighboring characters (Xue, 2003). In the 2000s, many researchers turned to conditional random fields or maximum entropy Markov models to tackle this task.

As the field progressed, approaches utilizing supervised or unsupervised features emerged (Zhao and Kit, 2008), contributing to the state-of-the-art performance in CWS. These techniques made significant strides in improving segmentation accuracy and efficiency.

A major turning point came around 2013 when researchers began incorporating neural networks into their work, revolutionizing the research landscape for CWS. Whether adopting feed-forward, recurrent, or convolutional neural network architectures, these approaches offered substantial benefits, particularly in reducing the need for labor-intensive data engineering tasks (Zheng et al., 2013; Pei et al., 2014; Chen et al., 2015).

Despite the promise of neural network-based methods, early trials did not consistently outperform non-neural systems. Refining these neural approaches became necessary to achieve their full potential in CWS tasks. However, over time, researchers made remarkable progress, and neural network-based methods eventually surpassed the capabilities of their non-neural counterparts, leading to significant advancements in the field.

In the realm of NLP, the emergence of transformers marked another significant milestone. Recent advancements in the field utilized a customized Transformer model for sequence tagging, resulting in the achievement of state-of-the-art performance (Duan and Zhao, 2020; Chou et al., 2023).

However, a common issue with many NLP models is that they are often trained and evaluated on contemporary datasets, limiting their applicability to historical or ancient texts. CWS is a field with a long history in Chinese NLP, resulting in the creation and continual evaluation of numerous datasets, like the ones listed below.

- SIGHAN Bakeoff 2005 (Emerson, 2005): It comprises diverse types of Chinese text, including news articles, fiction, and academic papers from various sources, such as CKIP (Academia Sinica) and City University of Hong Kong for traditional Chinese, and Beijing University and Microsoft Research (China) for simplified Chinese.
- Chinese Penn Treebank (CTB) [1]: It is one of the most widely used datasets for CWS research, and it exists in three versions: CTB6, CTB7, and CTB9.
- PKU Corpus (Yu et al., 2018): This dataset was collected from the People's Daily website and contains various text types, including news articles and editorials.
- Universal Dependency (UD) [2] for Mandarin Chinese: Similar to many other widely studied languages in NLP, Mandarin Chinese also has a Universal Dependency Annotation Scheme, which provides a standardized framework for dependency-based syntactic analysis.

Recently, to address the need for models capable of handling ancient Chinese texts, the EvaHan 2022 dataset [3] was created. It consists of annotated ancient Chinese text developed as part of the EvaHan project, a collaborative effort by several Chinese universities to build resources for ancient Chinese NLP. The Ancient Chinese language dates back to around 1000BC-221BC.

[1] https://www.cs.brandeis.edu/ clp/ctb/

[2] https://universaldependencies.org/treebanks/zh_pud

[3] https://circse.github.io/LT4HALA/2022/EvaHan

Thanks to the impressive performance of the models trained on these contemporary data, many off-the-shelf solutions are widely adopted in the context of digital humanities research, facilitating the exploration and analysis of Chinese texts from various time periods and genres:

- LAC (Jiao et al., 2018) is a joint lexical analysis tool developed by Baidu's NLP Department, which realizes the functions of Chinese lexical segmentation, lexical annotation, and proper name recognition.
- Jieba [4] is a module that is specifically used for CWS.
- Stanza (Qi et al., 2020) : This library was created by the Stanford NLP Group. It contains different tools for linguistic analysis such as POS tagging, lemmatization, segmentation and handles 66 languages including Chinese.
- SnowNLP [5] is a class library written in python inspired by TextBlob. It was created specifically to process Chinese.
- THULAC (Maosong Sun et al., 2016) is a set of Chinese lexical analysis toolkit developed and launched by the Laboratory of NLP and Social Humanities Computing of Tsinghua University, with the functions of Chinese lexical segmentation and lexical annotation.

These tools were designed with a strong emphasis on achieving high performance and user-friendliness. Nevertheless, it remains crucial to evaluate their performance when applied to specific datasets and Chinese language variants.

2 Dataset creation

Within the scope of our project, our primary objective was to evaluate various tokenizers on our specific dataset. The aim was twofold: first, to estimate the performance of off-the-shelf tokenizer tools, and second, to address the possibility that these readily available solutions might not yield suitable results for our unique data. In such a scenario, we planned to develop a custom model tailored to the requirements of our dataset.

In order to delve deeper into the development of the language during the particular period under examination, we engaged in manual tokenization of sentences extracted from the *Shenbao*. This involved carefully segmenting the text into individual words. We proceeded in two steps: for the first round, we aimed at annotating a large select of texts published between the years 1872 and 1947. Three annotators were trained on a sample of the selected corpus, after which they annotated separately 741 articles. For the second round, we selected 52 articles published between the years 1872 and 1907 that were annotated, curated, and used for evaluating our model.

By tokenizing these historical articles, we sought to gain relevant insights into the linguistic patterns, syntactic structures, and lexical variations present in the Chinese language during that specific historical timeframe. This approach allowed us to address the challenges posed by the absence of modern linguistic resources and the particularities that might arise in historical Chinese texts.

Through this comprehensive evaluation and analysis of word segmentation methods on our well-curated dataset, we anticipated obtaining a clearer picture of the tokenizer's efficacy in handling historical Chinese texts. Ultimately, this process played a vital role in shaping our subsequent decisions regarding the selection of the most appropriate tokenizer or the development of a custom model best suited to our research objectives and historical data.

To achieve this goal, we engaged two groups of three distinct annotators who were tasked with tokenizing the documents based on our guidelines. Each annotator was asked to add basic punctuation to mark the end of sentences. A blank space was introduced between tokens when their part of speech was different. Only place names, job titles, proper names of persons were left unspeparated.

Additionally, for the second round, a separate individual — a historian with high skills with classical and modern Chinese — was entrusted with curating the annotated data. This systematic approach ensured that the word segmentation process was carried out consistently and according to the specified guidelines, while also guaranteeing the quality and accuracy of the curated dataset.

Upon the successful completion of the annotation and curation processes, we obtained a carefully curated dataset, which we will call ENP-TOK [6], documented in Table 1. The agreement between annotators of the training set, calculated as the F1-score, is 75%, and the agreement between annotators of the evaluation set is 78%. These results

[4] https://github.com/fxsjy/jieba

[5] https://github.com/isnowfy/snownlp

[6] https://gitlab.com/enpchina/enp-tok

highlight the difficulty of this task during this period.

	Train	Eval
# Articles	741	52
# Sentences	11 132	396
# Characters	867 474	36 707
# Words	350 631	15 498

Table 1: ENP-TOK dataset statistics

ENP-TOK serves as a valuable resource for evaluating the performance and quality of various tokenizers, providing crucial observations for our research.

3 Experiments

By using this new annotated dataset, we have effectively assessed the performance of off-the-shelf tokenizers, particularly when applied to texts from the period of interest.

We conducted a thorough evaluation of the five most commonly used word segmentation tools. The results of this evaluation are summarized and presented in Table 2, providing measured references on the efficacy and suitability of each tokenizer for our specific historical texts.

Model	Precision	Recall	F-score
Jieba	42,94 %	53,61 %	47,61 %
Stanza	49.87 %	52.35 %	51.09 %
LAC	59.17 %	**69.09 %**	**63.74 %**
SnowNLP	**63.04 %**	52.45 %	57.26 %
thulac	47,24 %	61,42 %	53,41 %

Table 2: Off-the-shelf tools results on ENP-TOK dataset

In comparison to the performance achieved on contemporary datasets, the results obtained from these off-the-shelf models are significantly lower. It is imperative to verify and validate these models before employing them for larger-scale analyses to ensure the reliability of their outcomes. Although these off-the-shelf models offer convenience and quick implementation, we need to note that they may not represent the cutting-edge in terms of performance at the present moment.

To explore the potential for achieving better results without the need for additional annotation, we sought to evaluate more advanced and up-to-date models available in the HanLP library. HanLP is a NLP toolkit designed for production environments, focusing primarily on Chinese language processing. It provides pre-trained models specifically tailored for various datasets used in CWS evaluations.

Through the utilization of HanLP's (He and Choi, 2021) top-performing models for each dataset, we aimed to showcase the benefits of employing more recent and computationally heavier models, while emphasizing the crucial role of selecting the appropriate training set.

The outcomes of this experiment, detailing the performance of HanLP's best models, are documented in Table 3. These results highlight the potential advances that can be achieved in CWS using contemporary, state-of-the-art models without the need for additional manual annotations.

Based on these experiments, HanLP's models demonstrate significantly improved performance compared to off-the-shelf models in most cases. However, it is essential to acknowledge that the disparity in language models utilized—such as Electra, Roberta, and Bert—along with their distinctive learning regimes and training data, makes it challenging to predict which models would yield superior results on different types of datasets.

Model	Dataset	F-score
HanLP	SIGHAN2005	78.06 %
HanLP	CTB6	61.99 %
HanLP	PKU	63.68 %
HanLP	MSRA	76.91 %
HanLP	UD	68.94 %
FastHan	Multi	74.98 %

Table 3: Models results on ENP-TOK dataset

To address this challenge, FastHan (Geng et al., 2021) emphasizes the paramount importance of Generalization as a crucial attribute for any successful NLP toolkit. To achieve this, they developed a model trained on a diverse range of corpora specifically for CWS. The resulting model demonstrated improved performance across various sources.

In Table 3, we observe the results of applying this model to our dataset, confirming that it does outperform the majority of off-the-shelf tools. However, it falls slightly short compared to some models trained on more specific corpora.

Based on the findings, it becomes evident that employing corpora that align more closely with our specific requirements is the key to obtaining consistent and reliable results. Customizing the training data to suit the unique characteristics of our target text allows us to achieve optimal performance, ensuring that the model is better suited to handle the specificities of historical Chinese texts.

Given the lack of annotated data aligned with our specific period, it seemed wise to use at least annotated data from traditional Chinese or ancient Chinese.

4 Model training

To validate this hypothesis, we conducted extensive model training using three distinct datasets: CKIP, EvaHan, and ENP-TOK. Our aim was to assess the significance of both the training and inference datasets in shaping model performance. Beyond the influence of the CWS training data, we also delved into the impact of the language models employed to train these models.

All results presented in this section are averaged over 10 random initializations.

Initially, we used the BERT (Devlin et al., 2019) language model for Chinese. Because of its ability to deal with simplified and traditional Chinese (automatically translated), and its popularity.

Dataset	Precision	Recall	F-score
CKIP	71,25 %	78,75 %	74,81 %
EvaHan	78,12 %	78,19 %	78,16 %
ENP-TOK	**82.93 %**	**80.91 %**	**81.91 %**

Table 4: Result obtained on ENP-TOK when training from BERT-base-chinese[7], on several datasets

In view of the results presented in Table 4, the use of a training dataset aligned with our inference dataset yields better results, even if the quantity of data remains smaller than the other dataset. What is more, it seems that the Chinese we deal with is closer to ancient Chinese than to contemporary Chinese. Using data from EvaHan gives better results than CKIP.

Based on the results, presented depending on the quality of the original document in Table 4, using data from older Chinese sources leads to better outcomes. Specifically, utilizing data from EvaHan yields superior results. However, it is worth noting that the training of these models was initially initialized with word representations from contemporary language models. Therefore, to further investigate this, we repeated the previous experiment, but this time, we employed a language model trained on older Chinese data.

The findings suggest that incorporating data from historical Chinese sources can enhance the performance of the models for the task at hand. However, to fully explore the potential benefits, it is essential to consider the impact of using language models pre-trained on historical data (Wang and Ren, 2022), which may provide a more contextually relevant starting point for the training process.

Dataset	Precision	Recall	F-score
CKIP	65,68 %	76,62 %	70,73 %
EvaHan	83,55 %	81,11 %	82,31 %
ENP-TOK	**84.17 %**	**82.04 %**	**83.09 %**

Table 5: Result obtained on ENP-TOK when training from BERT-ancient-Chinese[8], on several datasets

The results, in Table 5, demonstrate that using language models aligned with the languages used in the sources leads to superior performance. These findings assert that the transitional Chinese that we are processing is more closely related to ancient Chinese than contemporary Chinese, which substantiates the need to train language models specifically on our data.

Additionally, it is interesting to examine the results obtained on CKIP. The misalignment between the language model and the annotated data negatively impacts performance when transferring to another dataset. This indicates that using a language model that is not well-aligned with the target data can lead to a decrease in performance when applying the model to different datasets. It highlights the importance of using language models that are tailored to the characteristics of the target data.

Aligning the language model with the linguistic properties of the data being processed, especially for historical languages like ancient Chinese, can significantly improve performance on various NLP tasks. Furthermore, understanding the impact of language model alignment and its effects on transfer learning can help researchers and practitioners make more informed decisions when deploying models across different datasets.

As of now, our efforts have culminated in the development of a model that achieves an impressive performance of over 83% . This achievement means a substantial leap in comparison to using the Jieba tokenizer, as our model reaches an F-score that is +35% higher.

While our results approach the levels attainable

[7] https://huggingface.co/bert-base-chinese

[8] https://huggingface.co/Jihuai/bert-ancient-chinese

on contemporary datasets, it is important to acknowledge that there are several potential avenues for further improvement.

The availability of this new annotated dataset opens up a wide array of possibilities for enhancing and fine-tuning the performance of the model in processing this type of data. We can pursue different paths, including linguistic analysis of the annotated dataset, or optimization at the training stage by incorporating specific language models.

With the aid of linguistic analysis, we can gain valuable insights into the unique characteristics and linguistic patterns present in modern historical Chinese texts. These results can help us refine the model to better capture and handle these nuances, ultimately improving its overall performance.

On the other hand, exploring various language models and implementing the most suitable one can significantly impact the model's ability to handle historical texts effectively. By selecting a specific language models that aligns closely with the linguistic traits of the historical period, we can boost its accuracy and adaptability to the complexities of the data.

This dataset is therefore used for continuous exploration and experimentation, empowering us to refine and optimize the model, or even pave the way for the development of advanced models tailormade for processing historical Chinese texts. This newfound dataset opens exciting possibilities for progress in this specialized domain.

5 Conclusion

Our study aimed to identify the most effective methods for CWS, specifically focusing on historical texts written in transitional Chinese. We created a novel, manually annotated dataset, derived from the *Shenbao*. This dataset provided us with a rich linguistic resource, allowing a comprehensive analysis of word segmentation methods for historical Chinese texts.

The evaluation of five commonly used off-the-shelf tokenizers revealed that while these models offer ease of use and quick implementation, their performance on our historical dataset was significantly lower than on contemporary datasets. This finding underscores the importance of validating models before using them for larger-scale analyses, to ensure the reliability of their outcomes. Further investigation with the HanLP library and the FastHan model demonstrated notable performance improvements, suggesting potential advancements in CWS using contemporary, state-of-the-art models without the need for additional manual annotations. However, the disparity between these models and the language models used for training makes it challenging to predict which models would yield superior results on different datasets.

Our exploration of different datasets and language models for training our segmentation model led us to a critical realization: using high-quality, relevant training data closely aligned with our target texts allowed us to achieve the best performance. This insight emphasized the importance of prioritizing data quality over quantity, and that customization of the training data to suit the unique characteristics of the target texts can lead to more accurate and reliable results.

Ultimately, our efforts culminated in the development of a model that achieved an impressive F-score of over 83% , a significant leap compared to using the Jieba tokenizer. While these results approach the levels attainable on contemporary datasets, there are several avenues for further improvement. The availability of ENP-TOK dataset opens up new possibilities for enhancing the model's performance, ranging from linguistic analysis of the annotated dataset to optimization at the training level by incorporating specific language models.

The improved CWS of transitional Chinese has significant implications for historians conducting research on China. The better the CWS, the more accurate and efficient the text analysis becomes. This improved accuracy can help streamline the research process, allowing historians to accurately segment and analyze large volumes of text, thus opening up new avenues of inquiry and enabling the exploration of previously unmanageable datasets. As most of the texts published in the period under consideration — including periodicals, books, diaries, etc. — were written in a style that closely matches the various forms on which we have trained our model, the tokenizer we propose can serve to unlock vast corpora of historical sources on which existing models fail. This can lead to more nuanced understandings of the wide range of historical texts made available through digitization, offering deeper insights into China's history.

This study illuminates the importance of dataset relevance and quality in achieving optimal results for CWS, particularly for transitional Chinese texts. The insights and methods derived from this study contribute significantly to the field of historical Chinese text analysis and provide valuable tools for historians working with these rich and complex linguistic resources.

Acknowledgment

- This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 788476).

- The authors wish to acknowledge fruitful discussions with Pierre Magistry. They also thank the annotators, Chang Yu-jun, Chiang Chia-wei, Hsu Wan-lin, and the curator Dr. Sun Huei-min (Institute of Modern History, Academia Sinica)

References

Bing Chen and Xiaoying Tai. 2009. A Hybrid Approach to Chinese Word Segmentation.

Haijing Chen. 2014. A Study of Japanese Loanwords in Chinese. Master's thesis, University of Oslo, Oslo.

Xinchi Chen, Xipeng Qiu, Chenxi Zhu, Pengfei Liu, and Xuanjing Huang. 2015. Long Short-Term Memory Neural Networks for Chinese Word Segmentation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1197–1206, Lisbon, Portugal. Association for Computational Linguistics.

Tzu Hsuan Chou, Chun-Yi Lin, and Hung-Yu Kao. 2023. Advancing Multi-Criteria Chinese Word Segmentation Through Criterion Classification and Denoising. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6460–6476, Toronto, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*, pages 4171–4186.

Sufeng Duan and Hai Zhao. 2020. Attention Is All You Need for Chinese Word Segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3862–3872, Online. Association for Computational Linguistics.

Thomas Emerson. 2005. The Second International Chinese Word Segmentation Bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Imre Galambos. 2021. Premodern Punctuation and Layout. In Jack W. Chen, editor, *Literary Information in China: A History*, pages 125–134. Columbia University Press.

Zhichao Geng, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2021. fasthan: A bert-based multi-task toolkit for chinese nlp. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 99–106.

John Christopher Hamm. 2021. Modern Punctuation and Layout. In Jack W. Chen, editor, *Literary Information in China: A History*, pages 135–142. Columbia University Press.

Xu Han, Hongsu Wang, Sanqian Zhang, Qunchao Fu, and Jun S. Liu. 2018. Sentence segmentation for classical chinese based on LSTM with radical embedding. *CoRR*, abs/1810.03479.

Han He and Jinho D. Choi. 2021. The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5555–5577, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Simon Hengchen, Ruben Ros, Jani Marjanen, and Mikko Tolonen. 2021. A data-driven approach to studying changing vocabularies in historical newspaper collections. *Digital Scholarship in the Humanities*, 36(Supplement_2):ii109–ii126.

Shilei Huang and Jiangqin Wu. 2018. A pragmatic approach for classical Chinese word segmentation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Zhenyu Jiao, Shuyu Sun, and Ke Sun. 2018. Chinese Lexical Analysis with Deep Bi-GRU-CRF Network. *ArXiv*.

Elisabeth Kaske. 2008. *The politics of language in Chinese education, 1895-1919*. Sinica Leidensia. Brill, Leiden. OCLC: 171268385.

Xiaoya Li, Yuxian Meng, Xiaofei Sun, Qinghong Han, Arianna Yuan, and Jiwei Li. 2019. Is Word Segmentation Necessary for Deep Learning of Chinese Representations? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3242–3252, Florence, Italy. Association for Computational Linguistics.

Pierre Magistry. 2021. Le(s)? chinois du Shun-pao .

Maosong Sun, Xinxiong Chen, Kaixu Zhang, Zhipeng Guo, and Zhiyuan Liu. 2016. THULAC: An Efficient Lexical Analyzer for Chinese. Original-date: 2016-05-17T05:05:05Z.

Barbara Mittler. 2004. *A newspaper for China?: power, identity, and change in Shanghai's news media, 1872-1912*. Number 226 in Harvard East Asian studies monographs. Harvard University Asia Center ; Distributed by Harvard University Press, Cambridge (Mass.).

Thomas S. Mullaney. 2017. Quote unquote language reform: New-style punctuation and the horizontalization of chinese. *Modern Chinese Literature and Culture*, 29(2):206–250.

Wenzhe Pei, Tao Ge, and Baobao Chang. 2014. Max-Margin Tensor Neural Network for Chinese Word Segmentation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 293–303, Baltimore, Maryland. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Binbin Wang. 1998. Gezai Zhongxi Zhijian de Riben—Xiandai Hanyu zhong de Riyu Wailaiy Wenti. ——"". Japan Exists between East and West— the Issue of Japanese Loanwords in Modern Chinese. . No.8. . *Shanghai Wenxue. . Shanghai Literature*, (8):71–80.

Pengyu Wang and Zhichen Ren. 2022. The Uncertainty-based Retrieval Framework for Ancient Chinese CWS and POS. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 164–168, Marseille, France. European Language Resources Association.

Nianwen Xue. 2003. Chinese Word Segmentation as Character Tagging. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 8, Number 1, February 2003: Special Issue on Word Formation and Chinese Language Processing*, pages 29–48.

Shiwen (Peking University) Yu, Huiming (Peking University) Duan, and Yunfang (Peking University) Wu. 2018. Corpus of Multi-level Processing for Modern Chinese.

Hai Zhao and Chunyu Kit. 2008. Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition. In *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing*.

Xiaoqing Zheng, Hanyang Chen, and Tianyu Xu. 2013. Deep Learning for Chinese Word Segmentation and POS Tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 647–657, Seattle, Washington, USA. Association for Computational Linguistics.

Introducing ChatGPT to a researcher’s toolkit: An empirical comparison between a rule-based and a large language model approach in the context of qualitative content analysis of political texts in Finnish

Ilona Kousa
University of Helsinki
ilona.kousa@helsinki.fi

Abstract

Large Language Models, such as ChatGPT, offer numerous possibilities and prospects for academic research. However, there has been a gap in empirical research regarding their utilisation as keyword extraction and classification tools in qualitative research; perspectives from the social sciences and humanities have been notably limited. Moreover, Finnish-language data have not been used in previous studies. In this article, I aim to address these gaps by providing insights into the utilisation of ChatGPT and drawing comparisons with a rule-based Natural Language Processing method called Etuma. I will focus on assessing the effectiveness of classification and the methods' adherence to scientific principles. The findings of the study indicate that the classic recall and precision trade-off applies to the methods: ChatGPT's precision is high, but its recall is comparatively low, while the results are the opposite for Etuma. I also discuss the implications of the results and outline ideas for leveraging the strengths of both methods in future studies.

1 Introduction

The field of Natural Language Processing (NLP) has recently undergone a significant transformation, largely driven by the widespread adoption and popularity of large language models (LLMs). LLMs, such as ChatGPT, offer numerous possibilities and prospects for academic research as well. Many researchers who have previously relied on traditional NLP methods are now considering the future trajectory of the field. The question arises: will other NLP methods become obsolete, with LLM applications replacing them in research?

Since the launch of ChatGPT in November 2022, there has been extensive discussion within the scientific community, and research articles have been published at an accelerated pace. Many of these studies have demonstrated that ChatGPT's performance in various tasks is comparable to that of humans in terms of quality.

In a study by Huang et al., (2023), ChatGPT was able to identify implicit hate speech well compared to humans. Guo et al., (2023) found that ChatGPT’s capabilities to answer questions from several domains including finance, medicine, law, and psychology, were on par with those of human experts. Gilardi et al., (2023) reported that ChatGPT even outperformed humans in annotation tasks including relevance, stance, topics, and frames detection. On the other hand, ChatGPT’s ability to produce consistent results has been questioned and caution has been advised regarding its application to text classification (Reiss, 2023). Some studies have found ChatGPT's zero-shot performance to be lacking, although prompt engineering and additional training have been shown to improve results (Shi et al., 2023; Yuan et al., 2023).

While ChatGPT has been extensively examined for a diverse range of tasks, there remains a gap in empirical research regarding its utilisation as a classification tool in qualitative research. Furthermore, perspectives from the social sciences and humanities have been notably limited thus far. In addition, Finnish-language data has not been used as research material.

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 102–113
December 1-3, 2023 ©2023 Association for Computational Linguistics

In this article, I will also introduce Etuma, an NLP tool that represents a traditional rule-based approach based on supervised learning methods, dictionaries, and grammar rules. The aim is to highlight the distinctive features of these two different approaches in one of the most common NLP tasks: text classification. I will focus on the qualitative content analysis of extensive datasets in the field of digital humanities, with a particular emphasis on topic classification, a central aspect of qualitative content analysis.

1.1 The scope of the study

The study aims to address the following research questions: 1) What distinguishes rule-based and LLM-produced classification in their effectiveness as qualitative content analysis tools? and 2) How viable are these methods in terms of scientific rigour, considering compliance with scientific principles such as reproducibility and transparency?

The motivation behind this research stems from a project that involves the categorisation of a large volume of Finnish-language texts. In the scope of this article, I will not discuss the underlying project and its results in detail, but rather focus on the description and validation of the methods. The main objective of this study is to describe the characteristics of different approaches and provide information to fellow researchers, who are considering using either method in qualitative content analysis.

In addition, beyond the features outlined in this article, there are several noteworthy concerns regarding the use of LLM techniques in research. These concerns include, for example, plagiarism and other unethical use, as well as challenges related to training data, including bias, misinformation, and vulnerability to adversarial attacks (Ray, 2023).

Firstly, I will describe the research setting, then report on the research materials, methods, and process. Then I will present and discuss the results and their limitations. I will conclude with insights and suggestions for future research.

2 Data and methods

The context of this study is an ongoing research project focusing on political energy discourse in Finland. In the project, my goal is to analyse the public political debate, specifically examining the comments made by citizens and politicians. The objective is to gain a comprehensive understanding of the issues underpinning the energy debate and to explore the various themes that emerge from the collected material. Given the large volume of the research material, I will employ a combination of automatic text analysis and qualitative methods (Guetterman et al. 2018; Jänicke et al. 2015; Grimmer & Stewart, 2013).

In research work, several important criteria should be considered when selecting a method or a research tool. The tool should be suitable for scientific research in general, adhering to rigorous standards of reliability, validity, and ethical considerations. In this case, an important feature was also the tool's proficiency in the Finnish language, to ensure its ability to process and analyse Finnish texts. Additionally, the tool needed to provide a comprehensive overview of large volumes of research material by effectively categorising it into relevant topics.

With these criteria in mind, I sought to explore whether the widely popular ChatGPT could be a suitable method for conducting the analysis required for the project. In my previous work, I have utilised the Etuma tool for keyword extraction and topic classification. Therefore, I decided to compare the two approaches to delineate the strengths and weaknesses of each method.

2.1 Data

The original corpus was collected as a part of the broader research project. It consists of 110,295 social media comments from August 2022 to August 2023 and 25,872 parliament speeches from February 2022 to March 2023. The social media comments were collected from a web scraping tool called Mohawk Analytics (Legentic 2023) and the transcribed parliament speeches were downloaded from a database known as Parliament Sampo (Hyvönen et al, 2022).

For the purposes of this article, I limited the material to a smaller subset so that it would be easier to qualitatively assess the analysis results produced by each method. I employed a keyword search (“electric car” AND “subsidy”; “sähköauto” AND “tuki” in Finnish) to filter texts discussing a specific topic of interest in the project: electric car subsidies offered by the Finnish government. The subset corpus comprised 40 social media comments and 33 parliamentary speeches.

The social media data included 21 tweets from Twitter (currently X), 19 online news comments, and 4 discussion forum posts. The parliament speech corpus consisted of 13 speeches from the Finns party, 3 speeches from the Social Democratic party, 3 speeches from the Centre party, 3 speeches from the Green party, and one speech each from the National Coalition party and the Christian Democrats. In addition, the material included 5 responses from government ministers from the Social Democratic party, the Centre party, and the Green party. The original language of the texts was Finnish, but keywords, topics, and text quotes have been translated into English for this article.

The social media comments were typically short, but their length varied between 20 and 155 words per comment. The comments were critical towards the research topic, as exemplified by statements such as "*Electric car subsidies go to the wealthy and electricity subsidies also benefit the wealthy. Because of the current government, we are all impoverished.*". Several comments included misspelled words.

The parliament speeches were more extensive, with their length varying between 72 and 662 words per speech. The speeches contained a considerable amount of specialised technical and administrative vocabulary, for example "*subsidies for the purchase of electric and gas cars and distribution infrastructure are necessary actions as we move towards a fossil-free transport system*" and did not contain much informal language, typos or misspellings.

I copied the original texts into an Excel file and recorded the analysis results obtained with different methods in their respective columns. To clean the data, I removed mentions targeted to specific users (identified with the '@' character) in social media comments. Additionally, I randomly selected a sample of 10 social media comments and 10 parliamentary speeches for validating the results. I will describe the categorisation and validation processes in more detail in the following sections.

2.2 Methods

In this study, I utilised both Etuma and ChatGPT4 for the identical task: extracting keywords from texts and categorising them into topics. I focused only on keyword and topic classification and excluded sentiment analysis. However, it's important to note that ChatGPT does not extract keywords in the traditional sense. Instead, it generates language based on the patterns it has learned from its training data.

The initial phase of the study started in a zero-shot setting, where no training data or pre-defined categories were used. I analysed the corpus in September 2023 using ChatGPT version 4, accessed through the Poe.com platform, and with Etuma's browser-based NLP tool. I will detail this process further in the subsequent section.

2.3 Research process

Figure 1 presents the key phases in the research process. During the study, I conducted both distant reading and traditional close reading in parallel (Jänicke et al., 2015). During the distant reading phase, I utilised computational methods to analyse the material based on topics and keywords, enabling a systematic examination of the data to identify patterns and trends. In the close reading phase, I engaged in an iterative process to uncover the topics present in the data, as well as their associated keywords and the context in which these keywords were discussed. This approach allowed for a deeper understanding of the data and its nuances. Finally, in the third step, the classification is refined to better align with the broader objectives of the research project, ensuring that it adequately captured the relevant information.

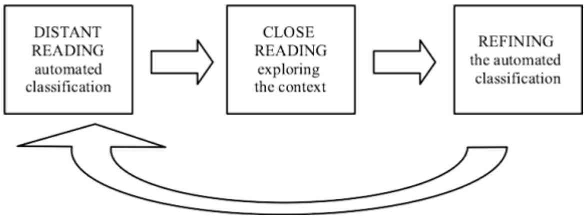

DISTANT READING automated classification
CLOSE READING exploring the context
REFINING the automated classification

Figure 1: Research process.

The approach of combining distant and close reading has been previously employed successfully. For example, Guetterman et al. (2018) conducted a study where they compared the results of qualitative analysis using three different methods: 1) close reading, 2) automatic text analysis, and 3) a combination of the two, by analysing the same materials in separate research groups. Their findings indicated that the combination of traditional close reading and automated distant reading yielded the most comprehensive, high-quality, and detailed results.

In the following sections, I will describe in more detail the distinctive features of the process for both methods separately.

2.3.1 ChatGPT

The term language model (LM) refers to systems that have been trained to predict the probability of a given token (character, word, or string) (Bender et al., 2021). ChatGPT has been pre-trained on large datasets consisting of web-crawled text, including conversations, and fine-tuned by humans with the Reinforcement Learning from Human Feedback (RLHF) method (OpenAI 2023a).

I employed ChatGPT 4 through the Poe.com platform. Users on this platform can create their own chatbot and customise its settings according to their preferences. This includes configuring a default prompt, which serves as the initial message for the chatbot, as well as setting a temperature value. Increasing the temperature parameter allows the predictive model to take more risks, suggesting less likely alternatives and thereby reducing result consistency (OpenAI 2023b).

The prompt plays a central role in determining what kind of results a ChatGPT-powered bot generates. After some testing with different prompt wordings and temperature values, I created a chatbot with the following prompt: “You are an advanced artificial intelligence for text analysis, and you need to classify given texts based on topics. One sentence can contain more than one topic. Extract as many topics as possible. The temperature setting is 0. Format the output to be a simple list of keywords that appear in the text and what topic the keywords are classified into.”.

The research process is illustrated in Figure 1. During the initial analysis phase, I input the texts individually into the same chat conversation and recorded ChatGPT's responses in an Excel table. In the zero-shot setting, the system autonomously identified 47 topics and 935 keywords within the data. Concurrently, I validated the classification by conducting a close reading of the original texts.

In the second analysis phase I experimented with a few-shot approach, providing more detailed instructions within the prompt about the specific topics I was interested in. I noticed that the more precise my requirements were defined, the better results I obtained. For instance, prompts like "extract relevant keywords and topics related to commuting" or "how are coronavirus aids and electric car subsidies linked in the texts" produced desired results but demanded accurate information or hypotheses about the material's content. In addition, ChatGPT's memory did not extend very far in the conversation, so it could not answer questions about the entire corpus.

In the third phase of the research process, I employed prompt engineering to refine the results and minimize the potential impact of a poorly formatted prompt on the outcomes by following the instructions of White et al., (2023). Among the four prompt enhancement strategies they proposed, I found "Question refinement" to align best with my needs, although in this case it did not lead to an improvement in recall. A report detailing example chat interactions of the prompt engineering experiment can be found in Appendix 2.

2.3.2 Etuma

Etuma’s technological foundation is rooted in NLP research conducted at the University of Helsinki (Lahtinen 2000; Tapanainen 1999), which has since been continued commercially by Etuma (Etuma 2023). Etuma performs several NLP tasks on texts, such as morphological, syntactic, semantic, and sentiment analysis.

A key function of Etuma is ontological classification, based on which it groups keywords referring to the same theme into more general classes called topics. For example, the keywords “*electric car”*, “*e-car”* and “*battery vehicle”* would be categorised into the same topic called Electric cars. It is important to note that although Etuma refers to the classification with the term topic, the method should not be confused with topic modelling methods, which are based on unsupervised machine learning, whereas Etuma employs dictionaries and supervised learning.

Using Etuma, I followed the same research process depicted in Figure 1. The initial analysis step involved uploading the original dataset in CSV format into the Etuma analysis system. Within the Etuma interface, I then applied filters as described above, to extract the specific sub-dataset relevant to this research. In the distant reading phase, the system identified 415 topics and 1621 keywords within the data. During the close reading phase, I conducted a review of the most frequently occurring topics and their corresponding keywords. Then I reviewed less frequent topics at a broader topic-name level.

In the subsequent phase, the emphasis is on refining the classification to improve the relevance and precision of the analysis by merging and splitting topics and transferring keywords between them. Etuma has a built-in user interface for these tasks, as refining the classification is an integral part of the research process. The extent of this phase depends on the goals of the research, the amount of material and precision of the classification. After the classification-validation process is completed, new classification rules are updated to the Etuma system, with the option that the customised rules can be reused. The purpose of the process is to improve the relevance of the classification to adapt to the specific requirements of the study. However, in this article I will focus on the zero-shot situation where no fine-tuning has been implemented.

3 Empirical analysis

In this section, I will present the key findings obtained from the analysis conducted by Etuma and ChatGPT on the corpora. Firstly, I will describe the characteristics of keyword extraction and topic classification for both methods, along with relevant examples. It is important to note that the purpose of these key figures is to compare the classifications, without taking a stance on what constitutes the ideal classification. Secondly, I will present a comparison of the methods using a smaller sample, employing traditional metrics such as recall, precision, and F1 score. This analysis will provide a quantitative evaluation of the performance of each method. Additionally, Appendix 1 contains a list of the most frequently occurring topics and keywords identified during the analysis.

3.1 Classification characteristics

Table 1 illustrates the differences in the number of unique keywords and topics identified by each method. As a general observation, ratio between the number of topics and keywords was similar in both corpora, indicating that the text type had no significant effect on the results.

		ChatGPT 4	**Etuma**
Social media	**Keywords**	246	435
	Topics	15	144
Parliament speeches	**Keywords**	722	1311
	Topics	40	378

Table 1: Unique keywords and topics in corpora

Keywords Both methods successfully analysed the Finnish-language material without significant deficiencies or shortcomings. However, there were differences in the keywords produced by the methods. The most noticeable difference was in the number of keywords: Etuma extracted more than one and a half times the number of unique keywords compared to ChatGPT. Additionally, Etuma tended to have more one-word keywords and ChatGPT generated more multi-word keywords.

The parliamentary speeches contained many acronyms. Both methods correctly classified common abbreviations such as EU (the European Union) and Yle (the Finnish Broadcasting Company). Etuma also extracted some acronyms from the parliamentary speeches (e.g., MAL, KAISU) but did not classify them to an exact topic. Initially, ChatGPT did not recognize these acronyms as keywords. When prompted separately, ChatGPT correctly classified MAL as "Maankäyttö, asuminen ja liikenne" (Land use, housing and transport) but did not identify "*KAISU*" as "Keskipitkän aikavälin ilmastopolitiikan suunnitelma" (Medium-term climate change policy plan).

ChatGPT correctly classified more names of Members of Parliament (e.g., Li, Tynkkynen) compared to Etuma. Typos and slang are common in social media materials. Etuma provides a list of keywords it does not recognize, and among them, there were 31 unique keywords that were misspelt and thus left uncategorized. Based on my observations, ChatGPT analysed typos correctly more frequently. However, a detailed analysis of the feature was not conducted in this study.

Topics In terms of unique topics, the difference between Etuma and ChatGPT was even more pronounced, almost tenfold. As can be deduced from the results, ChatGPT tended to employ broader topics (Economy, Politics), while Etuma's classification was more granular (Subsidies, Social security). Furthermore, it is worth noting that some of ChatGPT's unique topics overlapped (e.g., “Economics”, “Economics and Finance”, “Economy”, “Economy and Finance”), leading to even fewer distinct classification themes than the count of unique topic names identified.

Hallucination On a few occasions, ChatGPT demonstrated a behaviour known as hallucination, where it generated information that was not accurate or factual. For instance, it asserted that

"*Sulo Vileen*" (referring to a character from a Finnish TV series) is a colloquial term for ordinary Finns, akin to Joe Public in English. This occurrence also points to a limitation related to the training data in Finnish.

Prompting I tested various prompts with ChatGPT and repeated identical prompts in new chat interactions, which revealed that classification results for the same piece of text could change even though the content and prompt remained identical. As an example, during the initial analysis round, ChatGPT classified various keywords such as "*travelling to Spain*" "*musicians*" and "*price range*" under the same topic Social issues. However, in a new chat interaction, these same keywords were classified as International travel, Arts/Culture and Economy. This suggests that ChatGPT may have tried to simplify the classification by grouping less precise keywords into a smaller set of topics, indicating an internal learning mechanism guiding the classification.

3.2 Validation

To gain a more detailed understanding of the recall and precision levels of the methods, I conducted a comparative analysis with human classification. This involved calculating the traditional metrics of recall, precision, and the F1 score. During the validation phase, I randomly selected a sample of twenty texts from the material, consisting of ten social media posts and ten parliamentary speeches. Then I manually classified the texts by extracting the relevant keywords from them. At this stage, I tagged all potentially interesting keywords in the texts through which it would be possible to examine the material from various perspectives. Similarly, I did not provide specific instructions to Etuma and ChatGPT regarding the types of keywords to extract. As a result, I tagged a total of 151 keywords from the social media sample and 311 keywords from the parliament speech sample.

For each method, I compared the classification results with the human classification and calculated the recall using the following formula:

$$\frac{relevant\ extracted\ keywords}{all\ relevant\ keywords}$$

In addition, I calculated precision by reviewing the classification results and determining the number of keywords that were either left unclassified or classified incorrectly. The formula I used to calculate precision is as follows:

$$\frac{correctly\ classified\ keywords}{all\ extracted\ keywords}$$

The F1 score, a balanced measure that considers both precision and recall, is calculated as the harmonic mean of the two. I calculated the F1 score using the following formula:

$$2 * \frac{recall * precision}{recall + precision}$$

Table 2 presents the recall and precision levels, along with the F1 score that combines both metrics.

		ChatGPT 4	Etuma
Social media	**Precision**	0.96	0.70
	Recall	0.61	0.85
	F1 score	0.75	0.77
Parliament speeches	**Precision**	0.96	0.70
	Recall	0.58	0.81
	F1 score	0.72	0.75

Table 2 Recall, precision, and F1 score

Recall The recall level of Etuma's classification was higher in the social media sample (0.85) than in the parliamentary speech sample (0.81). For a single text, the recall ranged from 0.58 to 1.00, with an average of over 0.80 for both text samples. For ChatGPT the recall varied from 0.42 to 1.00 for individual texts, with an overall recall of 0.61 for the social media sample and 0.58 for the parliamentary speech sample.

A possible explanation for the difference between the two text types is that Etuma's tool is optimized for the analysis of relatively short customer feedback and survey responses, and not for the analysis of longer texts (Etuma 2023).

However, also ChatGPT's recall was higher for the social media sample. In the scope of this study, it is difficult to determine whether the difference is only due to the length of the texts, or whether the vocabulary and training materials used in the development of the methods also play a role. However, I observed that social media posts use more common language terms, while parliamentary speeches have more specialised terms that the tools did not always identify as keywords.

Precision Etuma's precision rate was 0.70 for both parliamentary speech texts and social media posts. However, different things affected the precision rate in the two samples. There were more misspelled words in the social media posts while there was more specialised vocabulary in parliamentary speeches. For example, from the sentence “*supplementary budget proposals allocate not only procurement support towards electric cars, but also support for ethanol and gas conversion opportunities”* Etuma did not recognize that the keyword “*gas conversion opportunities*” (kaasukonversiomahdollisuus in Finnish) referred to gas cars in this context.

ChatGPT's precision was high, 0.96 for both samples. Errors typically related to the interpretation of the correct topic, rather than to keyword extraction. For example, from the sentence in a social media post "*With this populist fake news, you can get a few votes in the elections, and nothing else*", ChatGPT classified the keyword "*elections*" (vaalit in Finnish) into a topic called Politics and the keyword "*votes*" (äänet in Finnish) into topic Social issues. In Etuma's analysis the precision rate was predominantly affected by uncategorised keywords. The results indicate that the precision of the results obtained is not significantly influenced by the type of text being analysed.

F1 score The F1 score, which takes into account both recall and precision, was slightly higher for Etuma in both the social media and parliament speech samples.

4 Discussion

In this section, I revisit the research questions I presented in the introduction. Firstly, I discuss the effectiveness of the classification in qualitative content analysis from the perspectives of key aspects such as recall, granularity, precision, and refinement. Secondly, I assess the alignment of the methods with scientific principles, specifically focusing on repeatability, transparency, and research integrity.

4.1 Effectiveness of classification

Recall In this study, Etuma's recall was higher compared to that of ChatGPT. The result reveals a fundamental difference between the approaches. While ChatGPT concentrates on summarising the content, Etuma aims to provide a comprehensive description of the content.

Granularity ChatGPT focused on the main points and tended to overlook rhetorical expressions and topics mentioned less frequently and more indirectly. A lack of detail was also observed in a previous study when comparing ChatGPT's responses to those of human experts (Guo et al., 2023). In situations where the corpus contains a significant amount of noise or irrelevant data, ChatGPT's ability to emphasise essential information can be beneficial. However, there are scenarios where researchers specifically seek nuanced details and rhetorical language, which may not align with ChatGPT's primary focus.

Precision As anticipated from prior research (Ortega-Martín et al., 2023), ChatGPT' performed well in semantic disambiguation and integrating cultural context into its classification. The adaptability of information related to cultural context stands out as notable strength of LLMs. Spelling mistakes and specialised vocabulary are more challenging for a dictionary-based approach because it is not feasible to add all possible spelling variants and special vocabulary to the ontology. Even though both methods are susceptible to the exclusion of specific terms, abbreviations, and misspelled words based on the vocabularies and training data utilised, this study revealed that ChatGPT outperformed Etuma in these regards.

In this study, there were no noticeable deficiencies in the knowledge of the Finnish language for either method. While I did not experiment with other languages, it is important to note that analysing a less common language like Finnish might not be as accurate or comprehensive due to the limited training material available.

Refinement A high recall or precision score does not automatically imply the relevance of results to the researcher. As I have described in this article, an important part of the research process is the validation and fine-tuning of the results in an iterative process. The workload involved in this step depends on the recall and precision of the initial analysis performed by the automated method. If the recall rate is high, it might be possible to enhance precision by refining the analysis. Conversely, if the precision rate is extremely low, the researcher faces a substantial workload in validating and fine-tuning the classification.

In this study, my attempts to fine-tune ChatGPT's results were not successful, as demonstrated in Appendix 2. However, if employed differently, it may be feasible to fine-tune ChatGPT's classification as well. On the other hand, Etuma has built-in tools designed to improve recall and precision as it is part of the method's standard process.

4.2 Compliance with scientific principles

Repeatability The methods differ in terms of reproducibility due to their distinct approaches. With the Etuma tool, the outcome of the analysis remains consistent, unless the researcher alters the classification rules. In contrast, a characteristic of ChatGPT is that identical input can yield different outputs. Moreover, during this study, I noticed that ChatGPT produced different results from the same text using the same prompt, a phenomenon that is in line with findings from earlier research (Ortega-Martín et al., 2023; Reiss, 2023).

In this regard, the method resembles qualitative analysis conducted by human analysts, as the classification performed by two different individuals may not be identical. A potential way to address this challenge could involve using similar approaches used to enhance the validity and reliability of human classification, like independently annotating the same material several times and then comparing the results.

Transparency With ChatGPT, transparency was impacted by the challenge of generating a manageable classification structure that could be easily documented and refined. ChatGPT operates as more of a black box, while Etuma offers greater transparency due to its classification being built upon predefined dictionaries.

The fine-tuning process of Etuma's classification is characterised by transparency and repeatability, as it is largely done manually, and every change leaves a trace in the system log. However, a challenge emerges from the extensive scope of classification, often requiring researchers to narrow their focus to, for example, a smaller subset of the corpus or the most prevalent topics.

Research integrity Despite the surrounding technological hype, researchers bear the responsibility to ensure that new technologies are not adopted too uncritically for scientific use. For example, various biases and information distortions due to training data and processes is an area that should be discussed. While this material appeared to be free from evident bias, it is important to acknowledge that in other types of content, biases may emerge. Additionally, ChatGPT's tendency to produce hallucinations, or inaccurate information, underscores the need for cautious evaluation of the data it generates. Furthermore, manually validating the analysis of a vast data set can be challenging, potentially allowing biases to go unnoticed.

In a broader perspective, it is important to consider the implications of tool development on the work of researchers. The findings of this study indicate that LLMs assume a significant portion of decision-making on behalf of researchers. While the idea of reducing workload is appealing, it is important to ensure that the autonomy of researchers is not compromised, potentially impacting the research process and even the results. As an example, an attempt to summarise complex information into broad topics may inadvertently overlook nuances or lead to potentially incorrect interpretations.

Moreover, relying solely on automated analysis tools can potentially direct researchers towards formulating research questions that align with the capabilities of the tools, rather than prioritising a comprehensive understanding of the phenomenon being studied. Additionally, it's important to note that although these tools are becoming more accessible, they do not always assure time savings or superior quality compared to manual methods. The utilisation of these tools can also be constrained by the fact that certain tools, such as public language model tools, may not be suitable for analysing sensitive data.

4.3 Strengths and limitations

The study has several strengths. Firstly, it addresses an existing knowledge gap by exploring the application of ChatGPT as a tool for qualitative analysis in Finnish. In addition, the perspective of the research is broadened using two distinct corpora. The study offers comparative insights for researchers who are considering employing either a large language model or a rule-based NLP approach for their analysis.

A limitation of the study is that the material is relatively narrow and focused on one specific research topic. Expanding the scope of the study would enhance the generalisability of the findings and provide a more comprehensive understanding of the methods' capabilities.

5 Conclusions

To summarise my findings, the utilisation of ChatGPT as a research tool poses challenges to the reliability of the research due to issues of repeatability and transparency. In the context of result usability, challenges arise from the occurrence of hallucinations and potentially from low recall.

A major limitation of a low recall rate is that it excessively restricts the researcher's autonomy in decision-making. In this project, my aim was to conduct a comprehensive classification that allows for a qualitative analysis of the material from various perspectives. Hence, I do not want the tool to determine what is important or interesting in the text on my behalf.

On the other hand, a drawback of the rule-based approach often lies in its lack of semantic meaning and context. Nevertheless, this deficiency can be addressed through refining, which, at least with the tools employed in this study, proves to be more straightforward with a rule-based method.

5.1 Implications for future research

In scientific research, repeatability and transparency are important features and the classification of qualitative content demands consistency, validity, and reliability. While ChatGPT may not yet substitute traditional NLP methods in these regards, it undeniably possesses strengths such as adept semantic analysis and information of cultural contexts.

In future research, it would be interesting to employ the methods in parallel and harness the strengths of both. Throughout the research process, I conceived numerous ideas on how to integrate the methods (indicated with a dashed line in Figure 2). The goal would be to utilise ChatGPT in a manner that ensures its shortcomings do not compromise the scientific principles.

Leveraging the LLM's capacity for semantic interpretations could enhance the semantic classification of another NLP method in the zero-shot phase, assisting in semantic filtering of research material based on the studied phenomenon.

In the close reading phase, LLM could aid researchers by generating automated summaries or in interpreting ambiguous or complex texts, suggesting alternative meanings and context to researchers in the validation process. The knowledge within the LLM based on the vast training data could extend beyond the corpus, aiding in the analysis of social discourse, for instance.

Furthermore, in the classification refinement phase, LLM's ability to identify semantic meanings and its creative capabilities could be used to formulate new topics or classification frameworks based on the feedback from the validation process.

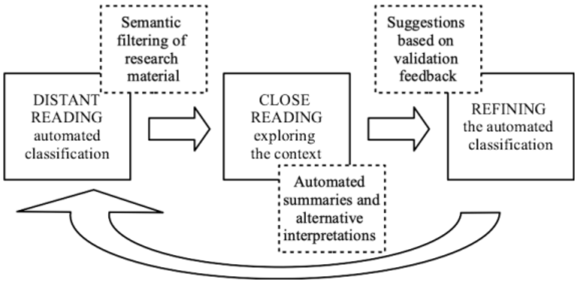

Semantic filtering of research material
Suggestions based on validation feedback
DISTANT READING automated classification
CLOSE READING exploring the context
REFINING the automated classification
Automated summaries and alternative interpretations

Figure 2: Research process combining rule-based and LLM approaches.

To address the question posed in the introduction about other NLP methods becoming obsolete, it is important to recognise that currently the principles of scientific research prevent ChatGPT from being a direct replacement for traditional NLP methods, at least in my research. However, its distinct advantages make it a potential complement to these methods, thereby enhancing my research toolkit.

Acknowledgments

I thank the Finnish Cultural Foundation, the Kone Foundation, and the Niilo Helander Foundation for supporting my research. I am also grateful to the anonymous reviewers and my colleagues for their comments on the manuscript.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Etuma (2023). *Etuma Natural Language Processing. Internal document.* Accessed on 26.9.2023.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). *ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks*. https://doi.org/10.48550/ARXIV.2303.15056

Grimmer, J., & Stewart, B. M. (2013). *Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Guetterman, T. C., Chang, T., DeJonckheere, M., Basu, T., Scruggs, E., & Vydiswaran, V. V. (2018). *Augmenting qualitative text analysis with natural language processing: methodological study. Journal of medical Internet research,* 20(6): e231.

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). *How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection.* https://doi.org/10.48550/ARXIV.2301.07597

Huang, F., Kwak, H., & An, J. (2023). *Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. Companion Proceedings of the ACM Web Conference* 2023, 294–297. https://doi.org/10.1145/3543873.3587368

Hyvönen, E., Sinikallio, L., Leskinen, P., La Mela, M., Tuominen, J., Elo, K., Drobac, S., Koho, M., Ikkala, E., Tamper, M., Leal R. and Kesäniemi J. (2022). Finnish Parliament on the Semantic Web: Using ParliamentSampo Data Service and Semantic Portal for Studying Political Culture and Language. *Digital Parliamentary data in Action (DiPaDA 2022), Workshop at the 6th Digital Humanities in Nordic and Baltic Countries Conference*, long paper, pp. 69-85, CEUR Workshop Proceedings, Vol. 3133, May, 2022.

Jänicke, S., Franzini, G., Cheema, M. F., & Scheuermann, G. (2015). *On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. Eurographics Conference on Visualization (EuroVis) - STARs,* 21 pages. https://doi.org/10.2312/EUROVISSTAR.20151113

Lahtinen, T. (2000). *Automatic indexing: an approach using an index term corpus and combining linguistic and statistical methods. Doctoral dissertation.* University of Helsinki.

Legentic (2023). *Legentic platform.* Accessed on 26.9.2023 at https://legentic.com/platform

OpenAI (2023a). *What is ChatGPT?* Accessed on 26.9.2023 at: https://help.openai.com/en/articles/6783457-what-is-chatgpt

OpenAI (2023b). *Quickstart. Adjust your settings.* Accessed on 25.9.2023 at: https://platform.openai.com/docs/quickstart/adjust-your-settings

Ortega-Martín, M., García-Sierra, Ó., Ardoiz, A., Álvarez, J., Armenteros, J. C., & Alonso, A. (2023). *Linguistic ambiguity analysis in ChatGPT.* https://doi.org/10.48550/ARXIV.2302.06426

Ray, P. P. (2023). *ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems,* 3, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Reiss, M. V. (2023). *Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark.* https://doi.org/10.48550/ARXIV.2304.11085

Shi, Y., Ma, H., Zhong, W., Tan, Q., Mai, G., Li, X., Liu, T., & Huang, J. (2023). *ChatGraph: Interpretable Text Classification by Converting ChatGPT Knowledge to Graphs.* https://doi.org/10.48550/ARXIV.2305.03513

Tapanainen, P. (1999). *Parsing in two frameworks: finite-state and functional dependency grammar. Doctoral dissertation.* University of Helsinki.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT.* https://doi.org/10.48550/ARXIV.2302.11382

Yuan, C., Xie, Q., & Ananiadou, S. (2023). *Zero-shot Temporal Relation Extraction with ChatGPT.* https://doi.org/10.48550/ARXIV.2304.05454

A Appendices

Appendix 1 Most frequent topics and keywords

Ranking	ChatGPT 4 topic and frequency	ChatGPT 4 keywords (translated into English)	Etuma topic and frequency	Etuma keywords (translated into English)
1.	Economy (n=53)	“wealthy”, “electric car subsidy”, “public spending”	Subsidies (n=72)	“electric car subsidy”, “coronavirus aid”
2.	Politics (n=53)	“government”, “left-wing”, “vote”	Cars (n=42)	“electric car”, “car”
3.	Social Issues (n=41)	“rural areas, “economic hardship”, “social and health services”	Government Organizations (n=34)	“government”, “EU”, “IMF”
4.	Environment (n=27)	“forest conservation”, “nature restoration”, “swamps	Fuel (n=20)	“fuel”, “gasoline”, “diesel”
5.	Automotive Industry (n=18)	“electric car subsidy”, “internal combustion engine”, “electric car”	Prices (n=19)	“price”, “expensive”, “cheap”
6.	Energy (n=14)	“electricity prices”, “energy prices”, “fossil fuels”	Purchasing (n=18)	“buy”
7.	Transportation (n=12)	“fuel taxation”, “electric car use”, “diesel cars”	Energy (n=15)	“electricity”, “energy”
8.	Economics (n=8)	“fuel tax”, “transportation costs”, “costs”	Payments (n=14)	“pay”
9.	Society (n=8)	“mail delivery”, “public transportation”, “accessible cars”	Taxes (n=12)	“taxes”
10.	Agriculture (n=7)	“agriculture”, “food producers”, “grain”	Income (n=11)	“low income”, “high income”
11.	Infrastructure (n=7)	“basic route maintenance”, “development project costs”	Climate Change (n=9)	“climate change”
12.	Economics and Finance (n=6)	“acquisition support”, “date of purchase”	Debt (n=9)	“debt”, “national debt”
13.	Industry (n=6)	“game companies”	Social Security (n=8)	“child benefit”, “state aid”
14.	Energy and Environment (n=5)	“climate actions”, energy self-sufficiency”, “energy crisis”	Wealthy (n=6)	“wealthy”
15.	Policy and Planning (n=5)	“support procedure”, “steering mechanisms”, “goals”	Saving (n=5)	“saving”

Appendix 2 Examples of prompt engineering attempts with ChatGPT 4 in September 2023

Chat interaction 1

Researcher	*“In the scope of text classification, suggest a better question that would allow me to get the most accurate classification of all possible keywords and topics that appear in the texts I provide.”*
ChatGPT 4	*“Given the following text, could you please identify and categorize all the keywords and topics present? Please provide a detailed breakdown of each keyword and topic and explain why each keyword or topic falls into its respective category based on the context of the text.”*
Researcher	*“Here is the text to be classified: [a social media comment from the corpus]”*

Result: One less keyword was extracted, and two additional topics were uncovered in the results following prompt engineering. Similar results were obtained with another social media comment and two parliamentary speeches.

Chat interaction 2

Researcher	*“In the scope of text classification, suggest a better question that would allow me to get the most accurate classification of all possible keywords and topics that appear in the texts I provide.”*
ChatGPT 4	*“Sure, here's a revised version of your question:* *"Given a set of texts, could you please perform a keyword extraction and topic classification? For each text, provide a list of extracted keywords along with their corresponding topic categories. The output should be formatted as a list where each entry includes the keyword and its associated topic category.""*
Researcher	*“Given a set of texts, could you please perform a keyword extraction and topic classification? For each text, provide a list of extracted keywords along with their corresponding topic categories. The output should be formatted as a list where each entry includes the keyword and its associated topic category.”*
ChatGPT	*“I'm sorry for any confusion, but as a text-based AI model, I don't have the capability to process a set of texts in real-time or perform keyword extraction and topic classification on a large dataset. However, I can certainly guide you on how you might approach this task using various natural language processing (NLP) techniques.* *Here's a general approach you might consider:* *Text Preprocessing: This involves cleaning the text by removing punctuation, converting all text to lowercase, removing stop words (common words like 'and', 'the', 'is', etc.), and stemming/lemmatization (reducing words to their root form).* *Keyword Extraction: There are several techniques for keyword extraction, including TF-IDF (Term Frequency-Inverse Document Frequency), TextRank, and RAKE (Rapid Automatic Keyword Extraction). These methods can help identify the most important words in each text.* *Topic Classification: Once you have the keywords, you can use them to classify the topics. One common approach is Latent Dirichlet Allocation (LDA), which is a type of probabilistic model that assumes each document is a mixture of a certain number of topics.* *The output of this process could be a list where each entry includes the keyword and its associated topic category. For example:* *- Keyword: 'climate', Topic: 'Environment'* *- Keyword: 'election', Topic: 'Politics'* *- Keyword: 'AI', Topic: 'Technology'* *Please note that this is a simplified explanation of the process and actual implementation might require a deeper understanding of NLP and machine learning techniques.”*

Result: No keywords or topics.

Chat interaction 3

Researcher	*“Given a set of texts, could you please perform a keyword extraction and topic classification? For each text, provide a list of extracted keywords along with their corresponding topic categories. The output should be formatted as a list where each entry includes the keyword and its associated topic category.”*
ChatGPT 4	*“Sure, I can certainly help with that. However, I need some example texts to work with. Could you please provide some?”*
Researcher	*“Here is the text: [a social media comment from the corpus]”*

Result: Two more keywords extracted, and two fewer topics compared to the previous classification.

Fly, Fly Little Comet!
Exploring Subtoken-Level Metaphorical Patterns in Finnish and Hungarian Texts. New Results from the FiHuCoMet Corpus.

Tímea Borbála Bajzát
Eötvös Loránd University, Doctoral School of Linguistics
bajzat.timi9696@gmail.com

Abstract

The FiHuCoMet Corpus was created to address the gap in the lack of a systematic comparison of metaphor research in Finnish and Hungarian (Bajzát and Simon, 2023). This study aims to: (i) expand the existing quasi-parallel corpus; (ii) explore subtoken-level metaphorical patterns comparatively in the examined languages with rich morphology. The analysis employs a MIPVU-inspired protocol for metaphor identification, the MetaID protocol (Simon et al., 2023). The sub-token level in this study refers to the morphological patterns that can be accessed at the subword level. Although this endeavor is not new, the comparative study conducted on a small-scale corpus has only revealed a few aspects of the potential of comparative metaphor analysis in the context of Finno-Ugric languages selected.

1 Introduction

A noticeable trend in recent years is the research on metaphorical structures, particularly from the perspective of cognitive semantics (Bolognesi and Werkman, 2023; Steen et al., 2010). This trend is evidenced by the significant efforts made over the past two decades to map metaphors as a linguistically accessible phenomenon with a comprehensive, language-specific focus (e.g., Huumo 2019; Máthé 2022). For instance, the development of language-specific adaptations of the MIPVU protocol (Steen et al., 2010), the most accurate and widely-used method for metaphor identification, has yielded numerous results examining the typological features of metaphorical elaborations (Nacey et al., 2019; Bogetić et al., 2019; Marhula and Rosiński, 2019; Urbonaité et al., 2019). However, the Uralic languages were not included in these efforts. This gap was identified by Bajzát and Simon (2023) when they introduced the theoretical and methodological framework of the FiHuCoMet project: the Finnish and Hungarian Comparative Metaphor Research Corpus based on quasi-parallel news texts. Their paper elaborates on the applicability of the adapted Hungarian, morpheme-based version of the MIPVU protocol (Simon et al., 2019, 2023) to the Finnish language. This morpheme-based process of MIPVU is equipped to address the metaphorical potentials that come from the typological features of agglutinative languages.

At the time of the FiHuCoMet project's inception, the corpus consisted of 5,116 tokens, allowing for only a small-scale analysis. Nonetheless, their results suggest relatively similar metaphorical linguistic elaborations in Hungarian and Finnish languages but reveal slight differences, such as variations in the frequencies of metaphorical expressions, metaphorical subtoken-level constructions, and the complexity of argument structures (Bajzát and Simon, 2023).

The method and preliminary results discussed above have inspired us to outline further research questions, which are the focal point of this study. Firstly, this paper introduces the expansion of the FiHuCoMet research corpus. The given study posits that the Hungarian and Finnish corpora exhibit similar metaphorical patterns at the subtoken level, considering their types and proportions.

2 Method

2.1 The Adapted MIPVU Method

The MIPVU method, adapted to Hungarian as the MetaID method, can be consistently applied to annotate metaphorical constructions in agglutinative languages (Simon et al., 2023) from a functional cognitive perspective (Langacker 2008). Due to space limitations, we cannot provide a detailed description of the adaptation process here, but we will highlight the most significant

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 114–119
December 1-3, 2023 ©2023 Association for Computational Linguistics

changes in the following paragraphs (for a thorough discussion, see Simon et al., 2023).

As metaphorization can occur at the subtoken level, the most notable change is that the annotation process is based on morphemes rather than words.

1. Viime kuu-**ssa** niitä oli
   last month-INE it-PART be-PST.3SG
   60. (Finnish subcorpus)
   60
   ('last month there were 60')

In the first Finnish example above (1), one can observe that time is conceptualized as a place by the highlighted inflectional suffix (*-ssA*, inessive case marking). The cited example represents a very conventional and grammaticalized way to express existence in time linguistically within the Finnish language. In many cases, the morphological units refer to a conceptualization that can be interpreted as an extension of the basic meaning based on similarity.

2. Jelenleg már 2012 halott-**ról**
   Currently already 2012 dead-DEL
   tudni. (Hungarian subcorpus)
   know-INF
   ('Currently 2012 deaths known')

In the second example (from the Hungarian subcorpus), we can observe a delative case referring not to the spatial position but to the topic of the process of knowing. The inflected noun (*halottról* 'about the dead') belongs to the infinitive (*tudni* 'know') as its argument, and the inflection is used as a case marker, which is obligatory in that specific construction ('know about something') (Sass et al., 2011, p. 171). Since the meaning associated with space is assumed to no longer be activated in such grammaticalized contexts, it is not marked as a metaphorical inflection (Simon et al., 2023). Steen et al. (2010) apply a similar method to handle highly grammaticalized elements.

The method does not attempt to detect the etymological aspects of metaphorization (Steen et al. 2010, p. 33–36). For example, in the context of compounds, the MetaID annotation schema relies on the principle of lexicalization, which is determined based on dictionaries (like the case of the Sesotho language, Seepheephe et al., 2019). If a compound word has not been lexicalized, it is not included as a unified entity in the dictionary we analyzed its component from the aspect of metaphorization. Moreover, only suffixes that do not change the word class of a given word form may receive tags, in line with the original MIPVU method.

Secondly, the modified annotation schema introduces a new set of tags to identify semantic relations based on cognitive grammar categories (Langacker 1987, 2013), with the aim of providing a more precise representation of metaphorical elaboration structures above the words. This approach allows us to detect extended patterns of metaphorization at the clause-level and facilitates cross-linguistic comparisons.

3. A terembem a **sötétség-et** csak
   The room-INE the darkness-ACC just
   a gyertyák és a mécsesek
   the candle-PL and the lantern-PL
   fény-e **tör**-te **meg**.
   light-POSS.3SG break-PST.3SG PREV
   ('The darkness in the room was broken only by the light of candles and lanterns)

The third example illustrates one instance of metaphorization in a multi-word expression. The verbal phrase (*törte* 'it broke') is annotated with the label of the metaphor-related expression because it initiates the metaphorical elaboration. At the same time, its arguments (*fénye* 'its light' and *sötétséget* 'darkness') also contribute to the metaphorization process, as we annotated them with the label of the metaphor-related argument (the full list of tags can be seen in Table 1).

The process of annotation is as follows: first, the text is split into morphological units, and then the basic and contextual meanings of the current morphological unit are determined using the dictionary, following the original MIPVU method. If an inter-domain mapping between the primary meaning and the contextual meaning can be assumed, the unit is marked as metaphorical. We annotate semantic relations separately and assess idiomaticity based on collocation (Simon et al. 2023).

The reliability of the MetaID procedure has been previously validated through assessments conducted on Hungarian language corpora (Simon et al., 2023). The kappa-values averaged 0.928 for mtags and 0.923 for mrel. Given that the overall performance of annotators surpasses the 0.8

threshold in kappa statistics (Carletta 1996, p. 252, Artstein–Poesio 2008, p. 22), the initial version of the adapted schema can be deemed reliable (Simon et al., 2023). It is essential to note that an inherent limitation in the current study lies in the absence of a comparable assessment for applying this procedure to the Finnish language yet. We intend to rectify this limitation in the upcoming research period.

2.2 The Brief Overview of the Tag Set

In the following tables (see Table 1 and Table 2) we attempt to introduce briefly the MetaID tag set and their semantics.

Tags	Function
MKK	Metaphor-related Expression
dMKK	Direct Metaphor-related Expression
MZ	Metaphor Flag
MKKimp	Metaphor-related Implicit Expression
MKI	Metaphor-related Inflection
MKA	Metaphor-related Argument
MKKomp	Metaphor-related Component
MKKid	Metaphor-related Idiomatic Expression
MKAid	Metaphor-related Idomatic Argument
MKKompid	Metaphor-related Idiomatic Component

Table 1: The tag set for identifying metaphorical structures.

Tags	Function
Tr (trajector)	It indicates the primary focal participant of the clause (Langacker 2008, p. 70–73)
Lm (landmark)	It signals the secondary focal participants of the scenario (Langacker 2008, p. 70–73)
Ela (elaboration)	Elaboration marks a non-specified elaborative operation
Poss (possessive)	This tag marks the possessive relation
Expm (explicating metaphorical meaning)	It signals the expressions used as a direct metaphor (MZ + dMKK).
R (unspecified semantic relation)	The label is used when two components of a multi-word unit move away from each other

Table 2: The relation set for identifying metaphorical structures.

3 The Project Infrastructure

3.1 The FiHuCoMet Research Corpus

In the process of building the FiHuCoMet research corpus, a fundamental organizational principle was followed: the incorporation of quasi-parallel texts in both its Hungarian and Finnish subcorpora. Here, 'quasi-parallel' does not mean processing identical source texts in both languages. Instead, it refers to processing texts with nearly identical content (report the same events), primarily comprising international political news obtained from online portals. These methodological choices were made to reduce potential biases in content and stylistic aspects, thus enhancing the objectivity of the studies (Bajzát and Simon, 2023). However, in the initial stage of corpus building, the subcorpora were relatively small, totaling 5,116 tokens (words) across both languages. The expanded corpus now contains 10,652 tokens. The principle of quasi-parallelism has been maintained during expansion. Nevertheless, thematically, the corpus has diversified and is no longer exclusively comprised of political news articles. The Hungarian-language news texts were drawn from Telex (Telex), while the Finnish-language news texts were collected from Helsingin Sanomat (Uutiset | HS.fi). The texts chosen for inclusion in the corpus were stored without their headlines and leads, as these structural elements are often duplicated within the main body of the text. Each of the Finnish and Hungarian subcorpora consists of 15 texts. As mentioned earlier, the sampling process was conducted in two stages. The first sampling took place in February 2022, while the second sampling was carried out in September 2023. The subcorpora can be categorized into the following major thematic units: international political news (F: 1,537 tokens; H: 3455 tokens), scientific and technological news (F: 1,114 tokens; H: 400 tokens), reports on natural disasters (F: 1,179 tokens; H: 932 tokens), news related to armed conflicts (F: 679 tokens; H: 679 tokens), and criminal news (F: 319 tokens; H: 358 tokens). As a result, the Hungarian subcorpus contains a total of 4,828 tokens, while the Finnish subcorpus comprises 5,824 tokens. Although the corpus of 10,652 words is relatively small for an extensive corpus linguistic study, the human capacity for manual annotation at this stage of the research did not allow for the processing of a larger sample. The present study provides exploratory feedback on the trends identified in the first phase of corpus building.

3.2 The Tools of Annotation

To annotate the Hungarian subcorpus, The Concise Dictionary of Hungarian (Pusztai ed. in chief, 2003) was employed. For the Finnish texts, The Dictionary of Contemporary Finnish (Kielitoimiston sanakirja) (Institute for the Languages of Finland 2022) was chosen to determine the default and contextual meanings of the expressions found. To measure idiomaticity in complex structures, a computational measuring tool was used, the word sketch browser of the Hungarian Web 2020 corpus (huTenTen20) and the Finnish Web 2014 corpus (fiTenTen), which provided association scores for collocations. Idiomaticity was evaluated using the LogDice typicality score (Rychlý, 2008), where a higher score (above 8.00) indicates a stronger association between the node and collocation candidates. In such cases, the method employed the MKKid tag to annotate metaphor-related idiomatic expressions and the MKKaid tag to denote their argument structure, or MKKompid when applicable (Bajzát and Simon, 2023, Simon et al., 2023).

The annotation was carried out using the WebAnno surface, designed by the CLARIN Research Infrastructure for Language Resources and Technology (Castilho et al., 2016). This platform allows for more transparent tagging of semantic relations and facilitates collaboration.

It's important to note that Hungarian texts were annotated by a native speaker, whereas Finnish texts were annotated by a non-native speaker with an upper-intermediate level of Finnish. Presenting this as a pilot study, we aim to inspire future collaborations between research communities, enabling cross-linguistic metaphor analysis with native speakers.

4 The Results

Figure 1 illustrates the proportion of tokens annotated with metaphorical expression labels across the entire corpus, with each column representing a news text. The most significant difference between Finnish and Hungarian news is noticeable in the 4th, 5th, 9th, 10th, 11th, and 12th pair of text. Generally, the extent of metaphorical marking in each subcorpus was similar in both Hungarian and Finnish. However, a noticeable difference in text length was observed in the case of two radically different text pairs (4th and 11th). In the fourth pair of texts, the Finnish text was relatively short (Finnish: 55 tokens; Hungarian: 211), while in the eleventh pair, the Hungarian text was significantly shorter than the Finnish one (Hungarian: 167 tokens, Finnish: 867 tokens). This confirms the observation that differences in text length can lead to a significant difference in their metaphorical markedness potential in 'quasi' parallel corpora.

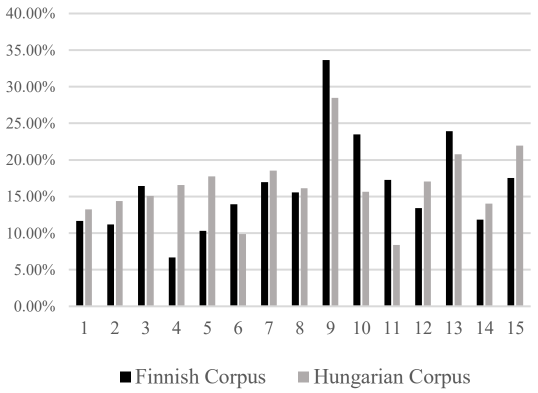


Figure 1: Relative frequencies of mtags in the subcorpora

For the other subcorpora, the length of the texts was relatively balanced. Furthermore, in nine subcorpora, the results indicate that Hungarian texts tend to have a slightly higher proportion of metaphorical tags compared to Finnish texts. The variance in sample sizes may lead to not only more frequent but also linguistically more complex metaphorical structures in Hungarian online news. These observations are in line with the results of the previous study (Bajzát and Simon, 2023). However, this can be nuanced by the fact that a higher proportion was also found in Finnish texts. Additionally, the slight differences can also be caused by the potential stylistic motivation. The higher occurrence and greater elaboration of metaphorical structures suggest that the speaker represented the events in a more sophisticated manner with greater stylistic potential.

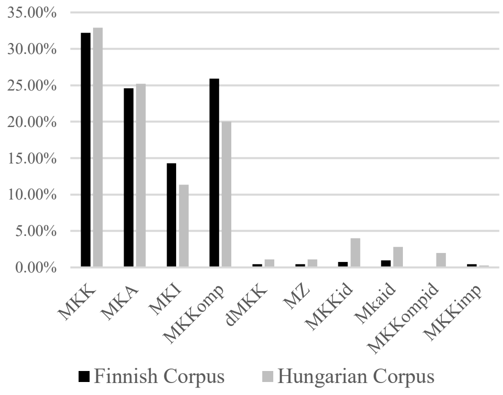


Figure 2: The proportions of mtags

Figure 2 illustrates the frequency of identified metaphorical units within the Finnish and Hungarian subcorpora. A notable difference is the higher prevalence of metaphorical elaboration at the morpheme level in the Finnish subcorpus, as indicated by the MKI bar. The varying proportions of MKKomp tags also demonstrate that the Finnish subcorpus exhibits more distinctive metaphorical elaborative operations. While the number of metaphorical expressions (MKK) and the examination of arguments do not show significant differences, it is observed that idiomaticity was more prevalent in Hungarian texts. This observation can further support the hypothesis of higher stylistic markedness in this subcorpus in terms of metaphorical constructions.

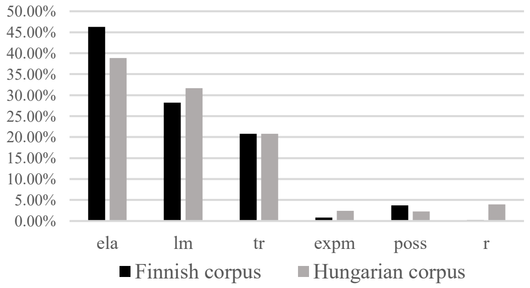


Figure 3: The proportions of mrel labels

Figure 3. illustrates the overall frequencies of labels assigned to the relations among metaphorical expressions in the Finnish and Hungarian subcorpora. The data reaffirm that the Finnish subcorpus has a higher proportion of metaphorical elaborative operations. A higher proportion of possessive metaphorical relations are also more characteristic of the Finnish material.

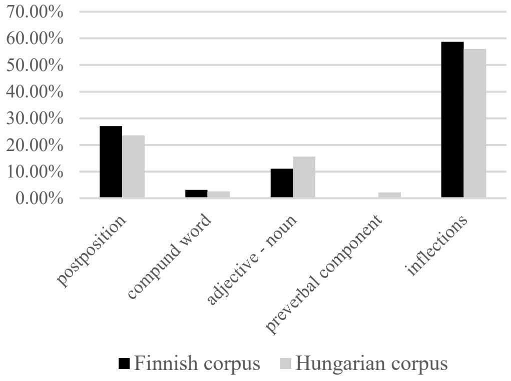


Figure 4: The distribution of elaborative relations

In Figure 4, we can observe slight differences in the distribution of semantic relations within metaphorical expressions, which highlight language-specific tendencies and patterns of morphological elaborative operations in metaphorization. In the Finnish subcorpus, a higher proportion of postpositions was measured (F: 27.11%; H: 23.62%) but the number of metaphorical adjective structures is lower in the Finnish texts (F: 11.12%; H: 15.62%). The inflections initiated the metaphorization are frequent in both languages.

5 Summary and Future Perspectives

The study aimed to report the latest findings from an ongoing project. Although the current FiHuCoMet corpus is still relatively small, it has more than doubled in annotated text volume compared to the previous phase. Recent results, particularly the analysis of subtoken-level metaphorization operations, confirm that while there are similarities in elaboration patterns between the two corpora, language-specific differences seem to be important as well. Looking ahead, it is justified to expand the corpus further and include texts of various types from multiple sources in parallel corpus building. Additionally, extending the metaphor identification map to include other Finno-Ugric languages is advisable for more comprehensive insights into comparative metaphor identification in these languages.

References

Ron Artstein and Massimo Poesio. 2008. *Inter-coder agreement for computational linguistics. Computational Linguistics 34:4*, pages 555–596. http://dx.doi.org/10.1162/coli.07-034-R2

Tímea B. Bajzát and Gábor Simon. 2023. Family relationship or family resemblance? A case study of comparative metaphor analysis in Finnish and Hungarian news texts. Under review.

Jean Carletta. 1996. *Assessing Agreement on Classification Tasks: The Kappa Statistic. Computational Linguistics 22:2*, pages 249–254.

Tuomas Huumo. 2019. Why monday is in front tuesday: On the uses of English and Finnish FRONT adpositions in SEQUENCE metaphors of time. *Linguistics 57:3*. pages 607–652. https://doi.org/10.1515/ling-2019-0010

Ksenija Bogetić, Andrijana Broćić, Katarina Rasulić. 2019. Linguistic metaphor identification in Serbian. In *Metaphor identification in multiple languages: MIPVU around the world*, John Benjamins, Amsterdam, pages 203–226. http://dx.doi.org/10.1075/celcr.22

Marianna, Bolognesi and Ana Werkman Horvat. 2023. *The metaphor compass. Directions for metaphor research in language, cognition, communication, and creativity.* Routledge, London, and New Work. https://doi.org/10.4324/9781003041221

Richard Eckart Castilho de, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevich, Anette Frank and Chris Biemann. 2016. A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84. https://www.aclweb.org/anthology/W16-4011

Kielitoimiston sanakirja. 2022. Kotimaisten kielten keskus, Helsinki, online publication. URN:NBN:fi:kotus-201433. https://www.kielitoimistonsanakirja.fi.

Langacker, Ronald W. 1987. *Foundations of cognitive grammar. Volume I theoretical prerequisites*, Stanford University Press, Stanford–California. https://doi.org/10.1515/9780804764469

Ronald W. Langacker. 2008. *Cognitive grammar: A basic introduction*. Oxford University Press, Oxford. DOI:10.1017/S0022226709005799

Langacker, Rondald W. 2013. Essentials of cognitive grammar. Oxford University Press, Oxford.

Joanna Marhula and Maciej Rosiński. 2019. Linguistic metaphor identification in Polish. In *Metaphor Identification in Multiple Languages: MIPVU around the world*, John Benjamins, Amsterdam, pages 183–202. http://dx.doi.org/10.1075/celcr.22

Zsuzsa Máthé. 2022. Space, time and transience. *Argumentum 18*, pages 273–286. 10.34103/ARGUMENTUM/2022/15

Susan Nacey, Aletta G. Dorst, Tina Krennmayr & Gudrun Reijnierse W. (eds.). 2019. *Metaphor identification in multiple languages: MIPVU around the world*. John Benjamins, Amsterdam, and Philadelphia. https://doi.org/10.1075/celcr.22

Ferenc Pusztai (eds.). 2003. *Magyar értelmező kéziszótár*, Akadémiai Kiadó, Budapest.

Pavel Rychlý. 2008. A lexicographer-friendly association score. In *Proceedings of Recent Advances in Slavonic Natural Language Processing. 6–9. Brno*. 13.pdf (muni.cz)

Nts'oeu Raphael Seepheephe, Beatrice Ekanjume-Ilongo, Motlalepula and Raphael Thuube. 2019. Linguistic metaphor identification in Sesotho. In *Metaphor identification in multiple languages: MIPVU around the world*, John Benjamins, Amsterdam, pages 267–287. http://dx.doi.org/10.1075/celcr.22

Bálint Sass., Tamás Váradi, Júlia Pajzs and Margit Kiss. 2011. Magyar igei szerkezetek. *A leggyakoribb vonzatok és szókapcsolatok tára*. Tinta Könyvkiadó, Budapest.

Gábor Simon, Tímea Bajzát, Júlia Ballagó, Zsuzsanna Havasi, Mira Roskó and Eszter Szlávich. 2019. Metaforaazonosítás magyar nyelv szövegekben: egy módszer adaptálásáról. *Magyar Nyelvőr 143: 2*, pages 223–247. http://nyelvor.c3.hu/period/1432/143208.pdf

Gábor Simon, Tímea B. Bajzát, Júlia Ballagó, Zsuzsanna Havasi, Emese K. Molnár and Eszter Szlávich. 2023. When MIPVU goes to No Man's Land: A New Language Resource for Hybrid, Morpheme-based Metaphor *Identification in Hungarian. Language Resources and Evaluation*. In press.

Gerard Steen, Aletta G. Dorst, Berenike Herrmann, Anna Kaal, Tina Krennmayr and Trijntje Pasma. 2010. *A Method for linguistic metaphor identification: From MIP to MIPVU*. John Benjamins, Amsterdam, John Benjamins. http://dx.doi.org/10.1075/celcr.14

Justina Urbonaitė, Inesa Šeškauskienė, Jurga Cibulskienė. 2019. Linguistic metaphor identification in Lithuanian. In *Metaphor identification in multiple languages: MIPVU around the world*, John Benjamins, Amsterdam, pages 159–181. http://dx.doi.org/10.1075/celcr.22

Machine Translation for Highly Low-Resource Language: A Case Study of Ainu, a Critically Endangered Indigenous Language in Northern Japan

So Miyagawa
National Institute for Japanese Language and Linguistics
Midoricho 10-2, Tachikawa, Tokyo
miyagawa.so.36u@kyoto-u.jp

Abstract

This paper explores the potential of Machine Translation (MT) in preserving and revitalizing Ainu, an indigenous language of Japan classified as critically endangered by UNESCO. Through leveraging Marian MT, an open-source Neural Machine Translation framework, this study addresses the challenging linguistic features of Ainu and the limitations of available resources. The research implemented a meticulous methodology involving rigorous preprocessing of data, prudent training of the model, and robust evaluation using the SacreBLEU metric. The findings underscore the system's efficacy, achieving a SacreBLEU score of 32.90 for Japanese to Ainu translation. This promising result highlights the capacity of MT systems to support language preservation and aligns with recent research emphasizing the potential of computational techniques for low-resource languages. The paper concludes by affirming the significant role of MT in the broader context of language preservation, serving as a crucial tool in the fight against language extinction. The study paves the way for future research to harness advanced MT techniques and develop more sophisticated models for endangered languages.

1 Introduction

The Ainu language, a polysynthetic and culturally rich language, has been traditionally spoken by the Ainu people in the northern regions of Japan, such as Hokkaido, Southern Sakhalin, and the Kuril Islands. Despite its intricate structure, the Ainu language faces significant endangerment. In 2009, UNESCO classified Ainu as a "critically endangered" language (Moseley, 2010), underscoring the critical need for efforts towards its preservation. The language's vulnerability is further highlighted by the dwindling number of native Ainu speakers and the loss of many Ainu dialects, including Sakhalin Ainu and Kuril Ainu.

The language itself has linguistic uniqueness such as polysynthesis and noun incorporation, which are characteristics of many indigenous languages in North America. Example 1 exemplifies its polysynthesis and noun incorporation.

(1) Hokkaido Ainu (Shibatani, 1990, 72)

Usa-opuspe
various-rumors
a-e-yay-ko-tuyma-si-ram-suy-pa
1SG-APL-REFL-APL-far-REFL-heart-sway-ITR

"I wonder about various rumors."[1]

Against this backdrop, this study aims to employ the advancements in Natural Language Processing (NLP) and Machine Translation (MT) to further our understanding and translation of the Ainu language. This research endeavors to leverage these technologies to contribute to the survival and revival of the Ainu language, especially given the urgency emphasized by its UNESCO status.

The ultimate objective of this study is to develop an AI-assisted educational program and a teaching robot to facilitate the learning and preservation of the Ainu language. The proposed program intends to incorporate several components like speech recognition, speech generation, part-of-speech tagging, and Universal Dependencies tagging, among other linguistic technologies, based on recent studies on the Ainu language.

Previous studies, such as the work of Nowakowski et al. (2019) on the Mingmatch—an n-gram model for Ainu word segmentation, and the creation of an Ainu folklore speech corpus by Matsuura et al. (2020b), the works mentioned in

[1] The list of abbreviations in the gloss: 1SG = first-person singular, APL = applicative, REFL = reflective, ITR = iterative.

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 120–124
December 1-3, 2023 ©2023 Association for Computational Linguistics

the next section, have laid the groundwork for this research. Building on these pivotal studies, this research aims to develop a robust NLP model that leverages the Marian MT as the primary translating model for Ainu to Japanese and Japanese to Ainu. The insights gained from this endeavor will inform the design of AI-assisted educational tools, thereby fostering the preservation and understanding of the Ainu language and culture.

2 Previous Literature

The potential for leveraging advanced computational techniques such as NLP and MT for language revitalization is gradually being explored. Previous work by Nowakowski et al. (2019) showcased a fast n-gram model for word segmentation of the Ainu language. This work signaled the potential of computational approaches for improving the accessibility and study of Ainu. Further efforts in this direction were made by Nowakowski (2020), who developed a digital corpus and core language technologies for Ainu. In another study, Nowakowski et al. (2017) proposed better text-processing tools for the Ainu language. These seminal works laid the groundwork for applying NLP techniques to Ainu, facilitating its digitalization.

Nowakowski's later work (Nowakowski et al., 2023) adapted a multilingual speech representation model for under-resourced languages through multilingual fine-tuning and continued pretraining. This showcased how techniques in NLP could be adjusted for low-resource languages like Ainu.

Efforts have also been made to apply speech recognition technology to the Ainu language. Matsuura et al. (2020b) developed a speech corpus of Ainu folklore and end-to-end speech recognition for the Ainu language. These authors also successfully utilized generative adversarial training data adaptation for very low-resource automatic speech recognition (Matsuura et al., 2020a). These studies significantly contribute to the field and provide a solid foundation for further exploration of NLP applications in low-resource languages. In terms of the linguistic study of Ainu, the work of Senuma and Aizawa (2017) in developing universal dependencies for Ainu and Ptaszynski et al. (2016) in improving part-of-speech tagging of the Ainu language have contributed significantly to the understanding of Ainu syntax and morphology, which is essential in developing accurate NLP tools.

The broader challenges of MT are aptly highlighted in the works of Koehn and Knowles (2017), which underscored the need for advanced techniques to address these challenges effectively. In language education, an innovative application of NLP tools was demonstrated by Nowakowski et al. (2020) through developing an Ainu language-speaking Pepper robot, indicating the potential of such technologies in promoting and preserving endangered languages. The insights and methodologies proposed in these studies pave the way for further exploration into using MT and other NLP technologies for language preservation, particularly for endangered languages such as Ainu.

3 Methodology

In this study, we utilized Marian MT. This efficient and adaptable open-source MT framework has demonstrated excellent performance in numerous research projects, particularly in scenarios involving low-resource languages (Ponti et al., 2021). Our choice for Marian MT was also informed by its inherent capacity to handle different language structures, an essential feature for polysynthetic languages such as Ainu (Ortega et al., 2020).

Our methodology commenced with data preprocessing obtained from multiple digital Ainu text sources. These included the Ainugo Archive from the National Ainu Museum[2], the Glossed Audio Corpus of Ainu Folklore from the National Institute for Japanese Language and Linguistics [3], and the ILCAA Ainu Language Resource from the Tokyo University of Foreign Studies[4]. We converted all the Katakana transcription into the Roman alphabet. The collected corpus was subsequently cleansed to eliminate redundancies and inconsistencies. Following this, we tokenized the data and segmented it into sentence pairs. Given the polysynthetic structure of Ainu (see Example 1), we took extra caution during the tokenization process to correctly separate individual morphemes. The number of sentence pairs of the Ainu original text and the Japanese translation is around 100,000.

We trained the Marian MT model with our prepared corpus, translating Ainu to Japanese and Japanese to Ainu directions. The model parameters were optimized through a learning rate schedule

[2] `https://ainugo.nam.go.jp/` (accessed June 24, 2023)

[3] `https://ainu.ninjal.ac.jp/folklore/` (accessed June 24, 2023)

[4] `https://ainugo.aa-ken.jp/` (accessed June 24, 2023)

combined with early stopping. The learning rate schedule gradually reduced the learning rate during the training process, thereby preventing the overfitting of the model to the training data. The early stopping technique mitigated overfitting by terminating the training when the model's performance on a validation set stopped improving (Almansor and Al-Ani, 2018).

We utilized the SacreBLEU metric to evaluate the performance of our MT system (Kim and Kim, 2022b). SacreBLEU provides a reliable and uniform method for comparing different MT systems or versions, implementing identical tokenization and detokenization procedures across all systems evaluated. It also accounts for multiple reference translations, thereby offering a more comprehensive evaluation of the translation quality (Kim and Kim, 2022a). This feature is especially beneficial for languages like Ainu, where the availability of parallel corpora is limited, and a given sentence could have multiple valid translations.

Our methodology thus encapsulated a combination of the Marian MT framework, rigorous preprocessing of the Ainu corpus, meticulous model training in Ainu to Japanese and Japanese to Ainu directions, and robust evaluation using the SacreBLEU metric. With this method, we developed a robust MT system capable of translating between Ainu and Japanese with significant accuracy.

4 Results

This chapter elucidates the outcomes of our MT experiments and provides an extensive discussion of their implications. The core of our study centered around two translation tasks: from Japanese to Ainu, from Ainu to Japanese, and both directions[5] (Table 1).

The model was trained on an extensive dataset gathered from various Ainu digital text sources for the Japanese-to-Ainu translation task. This resulted in a SacreBLEU score of 32.90, which implies a significant level of translation quality. This achievement showcases the model's ability to translate between these two disparate languages precisely. Notably, these results were obtained despite the inherent challenges posed by developing an MT system for a low-resource language like Ainu.

	Jpn.-Ain.	**Ain.-Jpn.**	**Bi-dir.**
Num. pairs	97,161	95,232	220,023
SacreBLEU	32.90	10.45	29.91

Table 1: Number of sentence pairs in used corpora and SacreBLEU scores of the best MT models in each case

The Ainu-to-Japanese translation task brought additional challenges, mainly due to the limited resources available for the Ainu language. Regardless, the MT system achieved a SacreBLEU score of 10.45. We also trained Marian MT for bidirectional translations, namely Japanese-Ainu and Ainu-Japanese, with a doubled corpus but reversed in the order of two languages in the latter half. The SacreBLEU score of this bi-directional experiment was 29.91, and the input can be both Japanese and Ainu, but the output is in the other language, which was not typed in the input.

Our research's relatively high SacreBLEU scores underline the feasibility of utilizing MT to aid language preservation and revitalization efforts. The results demonstrate that, even with limited resources, MT models can achieve a level of proficiency that renders them practical tools for Ainu learners and researchers (Kim and Kim, 2022b).

Additionally, our study supports the successes of previous attempts to apply computational techniques to the Ainu language. A notable instance is the Ainu speech recognition project by the Kawahara Lab at Kyoto University, whose results were documented in Matsuura et al. (2020b). Together, these studies underscore the potential contributions of NLP and MT technologies to preserve and revitalize endangered languages.

The outcomes of our study should inspire further exploration of MT applications in low-resource language contexts. Future endeavors could focus on refining the model's performance, expanding the dataset, and investigating how this technology can be integrated into interactive language learning platforms. Such efforts would further contribute to the revitalization of the Ainu language and culture.

5 Conclusions

This research project was undertaken to unlock the potential of MT in the preservation and revitalization of Ainu, a critically endangered and low-resource indigenous language of Japan. Grounded

[5] The models made in this study were published on HuggingFace. Ainu-to-Japanese: `https://huggingface.co/SoMiyagawa/ainu-2-japanese`, Japanese-to-Ainu: `https://huggingface.co/SoMiyagawa/japanese2ainu`, and bi-directional: `https://huggingface.co/SoMiyagawa/AinuTrans-2.0` (all accessed on June 24, 2023).

in the neural MT framework, Marian MT, and powered by a comprehensive dataset sourced from various Ainu digital text corpora, our study has made significant strides in demonstrating the feasibility and efficacy of MT in language preservation efforts.

The outcomes of our study are promising. With a SacreBLEU score of 32.90 for the Japanese to Ainu translation task, the quality of translation produced is commendable, particularly considering the challenging polysynthetic nature of the Ainu language (Ortega et al., 2020). Even more impressively, the model achieved a respectable SacreBLEU score of 29.91 for the Japanese-Ainu and Ainu-Japanese bi-directional translation task, underlining the robustness of the neural MT framework when dealing with complex, low-resource languages (Kim and Kim, 2022b).

These findings contribute to the expanding body of research that explores the potential of MT in bridging linguistic gaps and aiding in the preservation of endangered languages. Our results align with studies such as those conducted by Ranathunga et al. (2023), which emphasized the potential of neural MT for low-resource languages, and Kumar et al. (2021), which explored MT in low-resource language varieties.

Our research underscores the necessity for further work to leverage advanced MT techniques for low-resource languages, particularly where traditional linguistic databases may be limited or non-existent. By contributing to the intersection of Natural Language Processing (NLP), MT, and language preservation, our study offers a replicable methodology and highlights the importance of continuous innovation in these areas (Pilch et al., 2022).

Revitalizing endangered languages is complex and multifaceted, necessitating collaborative efforts across linguists, educators, technologists, and communities. Our research reiterates that MT and other language technologies are crucial in this process. While further refinement of models and expansion of datasets can enhance translation quality, our current findings underscore the significance of MT in the broader context of language preservation.

In conclusion, our study suggests that MT can make even the most resource-limited languages, like Ainu, more accessible. By facilitating communication, preserving cultural heritage, and fostering a deeper understanding of diverse human experiences, our research reaffirms the profound value of language preservation and the transformative power of technology in these endeavors. In light of UNESCO's classification of Ainu as "critically endangered," we believe our research can add a crucial layer of defense in the fight against language extinction and contribute to celebrating our shared linguistic heritage (Moseley, 2010).

References

Ebtesam H Almansor and Ahmed Al-Ani. 2018. A hybrid neural machine translation technique for translating low resource languages. In *Machine Learning and Data Mining in Pattern Recognition: 14th International Conference, MLDM 2018, New York, NY, USA, July 15-19, 2018, Proceedings, Part II 14*, pages 347–356. Springer.

Ahrii Kim and Jinhyeon Kim. 2022a. Vacillating human correlation of sacrebleu in unprotected languages. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 1–15.

Ahrii Kim and Jinhyun Kim. 2022b. Guidance to Pre-tokeniztion for SacreBLEU: Meta-Evaluation in Korean.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Sachin Kumar, Antonios Anastasopoulos, Shuly Wintner, and Yulia Tsvetkov. 2021. Machine translation into low-resource language varieties. *arXiv preprint arXiv:2106.06797*.

Kohei Matsuura, Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara. 2020a. Generative adversarial training data adaptation for very low-resource automatic speech recognition. *arXiv preprint arXiv:2005.09256*.

Kohei Matsuura, Sei Ueno, Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara. 2020b. Speech corpus of Ainu folklore and end-to-end speech recognition for Ainu language. *arXiv preprint arXiv:2002.06675*.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. Unesco.

Karol Nowakowski, Michal Ptaszynski, and Fumito Masui. 2017. Towards better text processing tools for the Ainu language. In *Language and Technology Conference*, pages 131–145. Springer.

Karol Nowakowski, Michal Ptaszynski, and Fumito Masui. 2019. Mingmatch—a fast n-gram model for word segmentation of the Ainu language. *Information*, 10(10):317.

Karol Nowakowski, Michal Ptaszynski, and Fumito Masui. 2020. Spicing up the game for underresourced language learning: Preliminary experiments with Ainu language-speaking Pepper robot. In *The 6st workshop on linguistic and cognitive approaches to dialog agents*.

Karol Nowakowski, Michal Ptaszynski, Kyoko Murasaki, and Jagna Nieuważny. 2023. Adapting multilingual speech representation model for a new, underresourced language through multilingual fine-tuning and continued pretraining. *Information Processing & Management*, 60(2):103148.

Karol Piotr Nowakowski. 2020. Development of a digital corpus and core language technologies for the Ainu language.

John E Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

Agnieszka Pilch, Ryszard Zygała, and Wiesława Gryncewicz. 2022. Quality assessment of translators using deep neural networks for polish-english and english-polish translation. In *2022 12th International Conference on Advanced Computer Information Technologies (ACIT)*, pages 227–230. IEEE.

Edoardo Maria Ponti, Julia Kreutzer, Ivan Vulić, and Siva Reddy. 2021. Modelling latent translations for cross-lingual transfer. *arXiv preprint arXiv:2107.11353*.

Michal Ptaszynski, Karol Nowakowski, Yoshio Momouchi, and Fumito Masui. 2016. Comparing multiple dictionaries to improve part-of-speech tagging of Ainu language. In *Proceedings of the 22nd Annual Meeting of The Association for Natural Language Processing, Sendai, Japan*, pages 7–11.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Hajime Senuma and Akiko Aizawa. 2017. Toward universal dependencies for Ainu. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 133–139.

Masayoshi Shibatani. 1990. *The languages of Japan*. Cambridge University Press.

Uncovering Gender Stereotypes in Video Game Character Designs: A Multi-Modal Analysis of Honor of Kings

Bingqing Liu*[1], Kyrie Zhixuan Zhou*[2], Danlei Zhu*[3], Jaihyun Park[2]
[1]Dalian Ocean University
liubingqing05@163.com
[2]University of Illinois at Urbana-Champaign
{zz78, jaihyun2}@illinois.edu
[3]BNU-HKBU United International College
r130201617@mail.uic.edu.cn

Abstract

In this paper, we conduct a comprehensive analysis of gender stereotypes in the character design of Honor of Kings, a popular multiplayer online battle arena (MOBA) game in China. We probe gender stereotypes through the lens of role assignments, visual designs, spoken lines, and background stories, combining qualitative analysis and text mining based on the moral foundation theory. Male heroes are commonly designed as masculine fighters with power and female heroes as feminine "ornaments" with ideal looks. We contribute with a culture-aware and multi-modal understanding of gender stereotypes in games, leveraging text-, visual-, and role-based evidence.

1 Introduction

Gender stereotypes, i.e., generalized preconceptions about characteristics or roles of a certain gender, broadly exist in video games, especially competitive games such as League of Legends (LoL) (Gao et al., 2017). Honor of Kings[1], the Chinese, mobile counterpart of LoL (Cheng et al., 2019), was released in 2015. It had over 145 million monthly active users in March 2022 (Wilson, 2022) and topped the global mobile game best-selling list with a revenue of 220 million dollars in July 2023 (Byshonkov, 2023). In this game, players can form their teams or randomly match teammates and opponents online. Teammates cooperate to grab resources, kill enemies, and ultimately, destroy the other team's base. Players can collect and play game characters/heroes who are categorized as warriors, assassins, mages, archers, tanks, or supports in the game.

Gender stereotypes in Honor of Kings have been studied through the lens of the female body (Zhang, 2022). Given the prominence of Honor of Kings in the Chinese video game industry, it potentially has wide-ranging impacts on people's conceptions of gender. In this work, we provided a more comprehensive understanding of gender stereotypes in the game character designs through the analyses of hero role assignments, the visual design of heroes and their skins, hero spoken lines, and hero background stories. We conducted a moral foundation analysis (Hopp et al., 2021) on the background stories to understand narratives around human morality, which often sees gendered expectations (Zhou et al., 2022a). We manually analyzed visual designs and hero lines. We also calculated descriptive statistics, i.e., percentages of male and female characters assigned to each role to understand the gender differences in hero role assignments.

We found that female heroes tended to be assigned to more traditionally feminine roles such as mages, while male heroes represented a wider range of roles such as warriors, tanks, and assassins. Female heroes were always designed with idealized looks and body shapes with revealing clothes, while not all male heroes exhibited an idealized appearance. In the spoken lines, male heroes were shaped as being eager to fight and protect, and having supreme power; female heroes were sentimental, objectified, and caring beauties. In the background stories, male heroes were more narrated on authority, while female heroes more on loyalty and sanctity.

Our inspection of gender stereotypes is multimodal, leveraging text-, visual-, and role-based evidence, and culture-aware, discussing gender roles in ancient Chinese culture. Based on the analysis, we propose future directions to mitigate gender stereotypes in video game character designs.

*The first three authors contributed equally to this paper.

[1] https://www.honorofkings.com/global-en/

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 125–131
December 1-3, 2023 ©2023 Association for Computational Linguistics

2 Related Works

The scholarship on games has drawn attention from a wide range of research communities. Some studies revealed the educational benefits of computer games (Mayer, 2019) and used games as an instrumental method for student engagement (Coller and Shernoff, 2009). Others focused on the collaborative behaviors of players, such as team formation (Gómez-Zará et al., 2019; Kim et al., 2017).

There has long been criticism about stereotyping and representation of gender in games. One of the early studies on gender stereotyping in video games was conducted on Super Mario Brothers, where the researcher argued that the narrative of the game could reinforce gender roles as players shared their identities as Mario characters (Sherman, 1997). Another study analyzed ten video games and argued that men were heavily over-represented in games as primary playable characters, power rested on male characters, and female characters remained supportive (e.g., nursing) (Friedberg, 2015).

At the visual level, Martins et al. found that female characters at low levels of photorealism were larger than the average American woman, while characters at the highest level of photorealism were thinner (Martins et al., 2009). The discrepancy between the real body shape of women and what was portrayed in games could lead to body dissatisfaction in women, attributing the media representation of a thin-ideal body (Grabe et al., 2008). Female character sexualization could lead to self-objectification of female players (Fox et al., 2015), low self-efficacy and self-worth (Behm-Morawitz and Mastro, 2009), and the acceptance of rape myths (Paul Stermer and Burkley, 2012).

Researchers have discussed stereotypes in Honor of Kings in terms of the distortion of historical facts. Yao and Chen found hero stories in this game have significantly reconstructed activity processes while largely preserving spatial circumstances, and partly fabricated social relationships among characters, resulting in the distortion of the historical timeline (Yao and Chen, 2022). Stereotyping and flattening of the hero images could affect the cultural image of the historical characters (Qiu, 2020).

So far, relatively few studies have focused on gender stereotypes in video games in China (Zhang, 2022; Sun, 2020; Chen, 2023). We enrich this literature with the current study.

3 Data Collection and Analysis

3.1 Data Collection

As of September 2023, there were 115 heroes in Honor of Kings. Among them, female heroes (N=36) accounted for 31%, male heroes (N=77) accounted for 67%, and heroes without a gender presentation (N=2) accounted for 2%. The skewed gender distribution of heroes already indicated a gender stereotype that men were more suitable and ready for "wars" or "battling" (Hutchings, 2008).

The *role assignment* and *background story* of heroes were collected from the official website of the game. The *hero lines* were collected from the in-game exhibition of heroes. If the lines of a certain hero were non-verbal, they were excluded from our analysis. The *visual analysis* involved both hero figures and their skins/outfits[2]. Both character lines and background stories are in Chinese.

3.2 Data Analysis

The fast-growing field of Natural Language Processing (NLP) was able in part due to existing datasets and models (Park and Jeoung, 2022) as well as metadata in digital archives (Dobreski et al., 2019). To take advantage of existing datasets and models and use them as an analytical lens, we adopted the Chinese Moral Foundation Dictionary (C-MFD) (Cheng and Zhang, 2023; Wu et al., 2019) to analyze the background stories of the heroes. C-MFD can be used for moral intuition detection and analysis in the Chinese language context. The creators of the dictionary drew on the Chinese translation of the English MFD (Hopp et al., 2021) and further fetched related words from an extensive Chinese dictionary based on Chinese moral concepts and word2vec. Categories in C-MFD include care vs. harm, authority vs. subversion, loyalty vs. betrayal, fairness vs. cheating, and sanctity vs. degradation, which are also present in MFD, as well as liberty vs. oppression, waste vs. efficiency, altruism vs. selfishness, diligence vs. laziness, resilience vs. weakness, and modesty vs. arrogance in the Chinese context.

We calculated moral foundation scores for each hero's *background story* in Chinese and compared the average scores for male and female heroes. One drawback of C-MFD is that it only provides the occurrence frequency for each moral dimension in the text without providing sentiment scores, which

[2] Each hero may have one or more skins; each skin may or may not include a new line.

prevents us from understanding if the narration of a certain gender leans toward the moral end or the immoral end of a moral dimension.

Since the *hero lines* were less rich in text with short lengths, there was not a sufficient overlap between them and C-MFD. Thus, we analyzed hero lines manually and used the translated lines to showcase our findings. Similarly, we manually analyzed the *visual features* of the heroes and their skins and made cross-gender comparisons. Two authors independently conducted the thematic analysis (Braun and Clarke, 2012) and regularly discussed to reach a consensus. We used a mind-mapping tool to organize the emerging themes and lines/visual features into a hierarchical structure.

Descriptive statistics were calculated to compare the hero *role assignments* across genders.

4 Results

4.1 Hero Role Assignment

There were only 36 female heroes in Honor of Kings compared to 77 male heroes. A closer look at the role assignments for different genders revealed that female heroes were mostly assigned as mages (44%), who tended to attack and control opponents from a distance. Fewer female heroes were assigned to roles known for hand-to-hand combat, such as warriors, assassins, and tanks. A large portion of male heroes were warriors (34%), and the remaining male heroes were distributed across other roles. More details are in Figure 1.

4.2 Visual Designs

Female heroes were designed with standardized physical features that conformed to traditional beauty standards and aesthetic preferences. They were presented as either beautiful or cute, with big breasts, a slim waist, and long legs. Nearly all female heroes had perfect faces that catered to traditional Asian aesthetics, evidenced by a pointy chin, big eyes, a high nasal bridge, a small mouth, and a perfect or even abnormal proportion (Zhang, 2012). In terms of dressing, we hardly saw female heroes wearing loose clothes. Even if a female hero was a warrior, the designers still intentionally exhibited the curve of her female figure with tight and revealing clothes. Such findings echoed prior studies which found women were portrayed and perceived as sex objects who embodied an idealized image of beauty (Dill and Thill, 2007).

The height or weight of male heroes varied, but most of them had abdominal muscles on their naked, upper bodies. Although most male heroes were designed to be tall, muscular, or robust to emphasize strength and fighting ability, not all male heroes were traditionally handsome – some of them were obese, had scars on their faces, or had other stereotypically imperfect characteristics. In general, less focus on the idealized image of beauty was put on male heroes than female heroes. A visual comparison between male and female characters can be seen in Figure 4 in the Appendix.

4.3 Hero Lines

Most lines, either those of male or female heroes, contained fighting-related elements, possibly due to the battling nature of the game. Yet, we still found differences between genders. The social identities of female heroes were limited to chefs, dancers, or goddesses, and they were often associated with purity and love. Male heroes had more social identities including warriors, princes, musicians, fortune tellers, and so on; they were often associated with conquering, defending, and fighting – traditionally masculine events.

4.3.1 Male Heroes

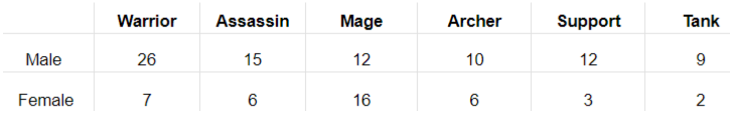
Fighters. Male heroes' lines were almost always about fighting, war, and violence, e.g., *"We fight for a common tomorrow," "Indomitable soul, inextinguishable fighting spirit, immortal heart," "War soul is not extinguishable," "I'm born for wars."* Such lines were also spoken with a firm and masculine tone.

Protecters. Male heroes often played the role of a protector for others, including their lovers, homes, and even Earth, e.g., *"Some people want to change the world, while others only want to protect their women," "Saint Seiya will always guard the love and peace of the earth," "In fairy tales, it is said that the prince overcomes thorns to find the imprisoned princess," "Eliminate evil relatives and keep the peace of the world."*

Suprememe Power. Male heroes' lines indicated the supreme power of men and their self-confidence (Meng and Literat, 2023), e.g., *"The devil is coming, like my miracle," "Telling you a secret, I'm invincible," "My only flaw is being too perfect."*

4.3.2 Female Heroes

Sentimental. Many female heroes' lines were about missing their lovers or other sentimental emo-



	Warrior	Assassin	Mage	Archer	Support	Tank
Male	26	15	12	10	12	9
Female	7	6	16	6	3	2

Figure 1: Role assignments for male and female heroes. One hero may have more than one role.

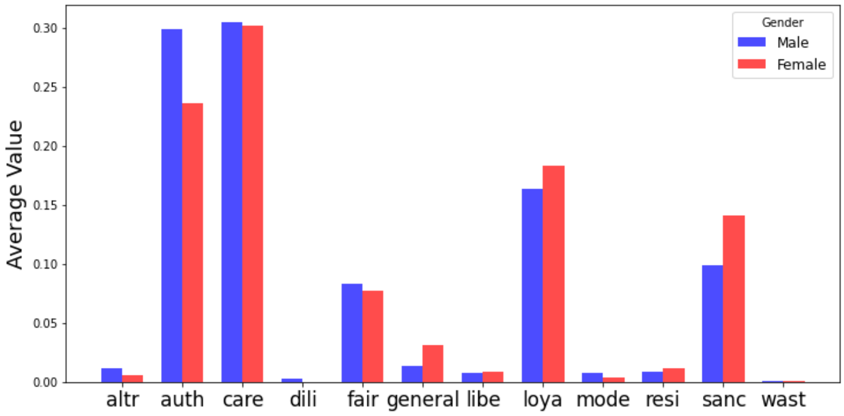


Figure 2: Moral foundation scores for male and female heroes in background stories. From left to right: altruism, authority, care, diligence, fairness, general, liberty, loyalty, modesty, resilience, sanctity, and waste.

tions, e.g., *"The east wind sends letters; the flower dynasty is as promised; we see each other every year, and we miss each other every year," "The saying goes that the magpies build a bridge over the cloud, and the destined one will run to you from the other end of the bridge."*

Serving and Caring. Female heroes often appeared as caring figures, such as a chef, a housewife, and a waitress. Example lines included *"I'm the one who cooks in the family," "I add sugar to the memory; guests, please taste it with heart."* Such lines were uttered in soft, gentle tones by the characters. On the contrary, when a male hero appeared as someone with cooking skills, their lines emphasized the food itself, e.g., *"Only love, justice, and food cannot be disappointed," "Hot and spicy from the depths of the soul."*

Appearance. Even if a female hero was designed as a warrior, her lines were still about the "ornamental" role of women instead of the fighter role, e.g., *"Reap your heart," "Acting as roses," "Yes, I'm tempting you."* Some lines were explicitly about appearance, e.g., *"I'm so cool and beautiful," "Beautiful girls never look back at explosions."*

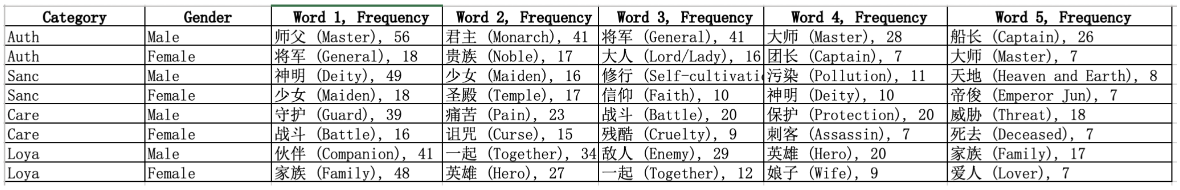
Objectified. Some female heroes' lines exhibited objectification of women, treating women as inferior people or objects, e.g., *"I'm your Christmas present tonight."* One female hero was designed as a dancer and addressed herself as a "concubine," which was a self-designation in ancient China where women were regarded as possessions of men with a lower social status.

4.4 Hero Background Stories

We compared occurrences of moral words between genders and identified notable differences in the authority/subversion, loyalty/betrayal, and sanctity/degradation moral dimensions (see Figure 2 for a full comparison). Even in moral dimensions with similar occurrence frequency for male and female characters (e.g., care/harm), the gendered narration of heroes was obvious. We further identified the top five common moral words related to these moral dimensions for both genders (Figure 3).

Authority/Subversion. Authority/subversion-related moral words were more often seen in the narrative of male heroes than in female heroes. A closer look revealed that male heroes were often narrated with words indicating positions of high authority such as "master," "monarch," "general," and "captain." e.g., *"This is the true face of Master Lu Ban and his genius creation, Lu Ban No. 7!"* Female heroes were less frequently described as a "general" or a "noble."

Loyalty/Betrayal. Female heroes were more narrated with loyalty/betrayal-related words than male heroes, such as "family," "wife," and "lover," e.g., *"The mission of the family and the responsibilities of the eldest sister fall upon her."* Male heroes were more linked with such words as "companion" and "enemy," again demonstrating their roles as



Category	Gender	Word 1，Frequency	Word 2，Frequency	Word 3，Frequency	Word 4，Frequency	Word 5，Frequency
Auth	Male	师父（Master），56	君主（Monarch），41	将军（General），41	大师（Master），28	船长（Captain），26
Auth	Female	将军（General），18	贵族（Noble），17	大人（Lord/Lady），16	团长（Captain），7	大师（Master），7
Sanc	Male	神明（Deity），49	少女（Maiden），16	修行（Self-cultivati	污染（Pollution），11	天地（Heaven and Earth），8
Sanc	Female	少女（Maiden），18	圣殿（Temple），17	信仰（Faith），10	神明（Deity），10	帝俊（Emperor Jun），7
Care	Male	守护（Guard），39	痛苦（Pain），23	战斗（Battle），20	保护（Protection），20	威胁（Threat），18
Care	Female	战斗（Battle），16	诅咒（Curse），15	残酷（Cruelty），9	刺客（Assassin），7	死去（Deceased），7
Loya	Male	伙伴（Companion），41	一起（Together），34	敌人（Enemy），29	英雄（Hero），20	家族（Family），17
Loya	Female	家族（Family），48	英雄（Hero），27	一起（Together），12	娘子（Wife），9	爱人（Lover），7

Figure 3: Top 5 moral words for male and female heroes in background stories. We only list top words in moral dimensions with relatively more occurrences.

fighters, e.g., *"Every night, he finds himself surrounded by thousands of enemies in his dreams."*

Sanctity/Degradation. Female heroes were more frequently linked with moral words about sanctity, such as "maiden" and "temple," emphasizing the purity and sanctity expectations of women, e.g., *"The maiden feels anger and pain for the unfair treatment."* Male heroes were also frequently linked with sanctity-related words, such as "deity" and "maiden," yet these words suggested their high power, e.g., *"Possess the powerful force of a deity."*

Care/Harm. Both genders were frequently associated with care/harm-related words, given the battling (harm) nature of this game. Commonly seen in male heroes' stories were "guard," "pain," "battle," "protection," and "threat," showing both tendencies of destruction and protection. Overall, male heroes were narrated as brave and violent, e.g., *"Bajie bravely charged to the forefront of the team."* On the other hand, female heroes were associated with less powerful words such as "curse" and "deceased," e.g.,*"The supreme empress of Chang'an City will never forget her cursed destiny and the dream of an ideal kingdom."*

5 Discussion and Future Work

By presenting a comprehensive analysis of visual designs, role assignments, spoken lines, and background stories, we uncovered the prevalent gender stereotypes in Honor of Kings. Our role, text, and visual analyses echoed each other, depicting how the game character design consistently reinforced gender stereotypes. The idealized looks of female characters, combined with the single aesthetic perspective of women in media, i.e., being slim, fair skin, etc. (Grabe et al., 2008), may create pressure and anxiety for women.

In the game, male characters are designed as people in power, fighters, and decision-makers, extending the traditionally perceived role of men in Chinese society (Xie, 2013). Female heroes are designed as feminine ornaments with ideal looks. They tend to play supportive roles (e.g., mage) in battles. Such findings echo prior studies on stereotypes in games (Grabe et al., 2008; Martins et al., 2009; Friedberg, 2015). Female characters are also stereotypically shaped as emotional and without decisive, competitive, and strong traits.

In traditional/ancient Chinese culture, male superiority and female inferiority were simultaneously emphasized – men tended to dominate economically and socially, while women were expected to be responsible for childcare and household chores in the family (Zhou et al., 2022b, 2023b; Zhou and Sanfilippo, 2023). The traditional division of gender roles in/outside the family still exists nowadays, though women are increasingly pursuing careers and independence (Gui, 2020). Such power imbalance and stereotypical gender narration are reflected in Honor of Kings. While prior research criticized the distortion of history in this game (Yao and Chen, 2022; Qiu, 2020), the gender dynamics are sarcastically true to reality.

With this exploratory uncovering of gender stereotypes in a Chinese video game, we aim to spur more research in this relatively under-investigated cultural context. We suggest several lines of research for future investigation of stereotypes in video game design in specific cultural contexts. First, user studies on how Chinese game players perceive such gender stereotypes are encouraged, as most prior studies were conducted in Western contexts. The persisting, traditional gender roles in East Asian society may lead to either people's desensitization to stereotyping or stronger resistance to it (Lee, 2017). Second, the video game industry has been long known as a regime of masculine domination (Styhre et al., 2018; Dunlop, 2007). Including more women and gender minorities in the game design and development lifecycle, as well as providing educational interventions to equip designers and developers with gender awareness, are key to mitigating gender stereotypes in games (Zhou et al., 2023a).

Acknowledgments

We sincerely thank the anonymous reviewers for their supportive and constructive feedback, which we have leveraged to polish up the paper.

References

Elizabeth Behm-Morawitz and Dana Mastro. 2009. The effects of the sexualization of female video game characters on gender stereotyping and female self-concept. *Sex roles*, 61:808–823.

Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

Dmitriy Byshonkov. 2023. Sensor tower: Top 10 mobile games by revenue and downloads in july 2023. *GameDev Reports*.

Dongna Chen. 2023. Female characters' images in chinese otome game and woman stereotype.

Calvin Yixiang Cheng and Weiyu Zhang. 2023. C-mfd 2.0: Developing a chinese moral foundation dictionary. *Computational Communication Research*, 5(2):1.

Ziqiang Cheng, Yang Yang, Chenhao Tan, Denny Cheng, Alex Cheng, and Yueting Zhuang. 2019. What makes a good team? a large-scale study on the effect of team composition in honor of kings. In *The World Wide Web Conference*, pages 2666–2672.

Brianno D Coller and David J Shernoff. 2009. Video game-based education in mechanical engineering: A look at student engagement. *International Journal of Engineering Education*, 25(2):308.

Karen E Dill and Kathryn P Thill. 2007. Video game characters and the socialization of gender roles: Young people's perceptions mirror sexist media depictions. *Sex roles*, 57(11-12):851–864.

Brian Dobreski, Jaihyun Park, Alicia Leathers, and Jian Qin. 2019. Remodeling archival metadata descriptions for linked archives. In *International Conference on Dublin Core and Metadata Applications*, pages 1–11.

Janet C Dunlop. 2007. The us video game industry: Analyzing representation of gender and race. *International Journal of Technology and Human Interaction (IJTHI)*, 3(2):96–109.

Jesse Fox, Rachel A Ralston, Cody K Cooper, and Kaitlyn A Jones. 2015. Sexualized avatars lead to women's self-objectification and acceptance of rape myths. *Psychology of Women Quarterly*, 39(3):349–362.

Jared Friedberg. 2015. Gender games: A content analysis of gender portrayals in modern, narrative video games.

Gege Gao, Aehong Min, and Patrick C Shih. 2017. Gendered design bias: gender differences of in-game character choice and playing style in league of legends. In *Proceedings of the 29th Australian Conference on Computer-Human Interaction*, pages 307–317.

Diego Gómez-Zará, Matthew Paras, Marlon Twyman, Jacqueline N Lane, Leslie A DeChurch, and Noshir S Contractor. 2019. Who would you like to work with? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–15.

Shelly Grabe, L Monique Ward, and Janet Shibley Hyde. 2008. The role of the media in body image concerns among women: a meta-analysis of experimental and correlational studies. *Psychological bulletin*, 134(3):460.

Tianhan Gui. 2020. "leftover women" or single by choice: Gender role negotiation of single professional women in contemporary china. *Journal of Family Issues*, 41(11):1956–1978.

Frederic R Hopp, Jacob T Fisher, Devin Cornell, Richard Huskey, and René Weber. 2021. The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior research methods*, 53:232–246.

Kimberly Hutchings. 2008. Making sense of masculinity and war. *Men and Masculinities*, 10(4):389–404.

Young Ji Kim, David Engel, Anita Williams Woolley, Jeffrey Yu-Ting Lin, Naomi McArthur, and Thomas W Malone. 2017. What makes a strong team? using collective intelligence to predict team performance in league of legends. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 2316–2329.

Yean-Ju Lee. 2017. Multiple dimensions of gender-role attitudes: Diverse patterns among four east-asian societies. *Family, Work and Wellbeing in Asia*, pages 67–87.

Nicole Martins, Dmitri C Williams, Kristen Harrison, and Rabindra A Ratan. 2009. A content analysis of female body imagery in video games. *Sex roles*, 61:824–836.

Richard E Mayer. 2019. Computer games in education. *Annual review of psychology*, 70:531–549.

Xingyuan Meng and Ioana Literat. 2023. # averageyetconfidentmen: Chinese stand-up comedy and feminist discourse on douyin. *Feminist Media Studies*, pages 1–17.

Jaihyun Park and Sullam Jeoung. 2022. Raison d'être of the benchmark dataset: A survey of current practices of benchmark dataset sharing platforms. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 1–10.

S Paul Stermer and Melissa Burkley. 2012. Xbox or sexbox? an examination of sexualized content in video games. *Social and Personality Psychology Compass*, 6(7):525–535.

Zifan Qiu. 2020. Stereotyped and flattened: the characteristics and cultural influence of hero reconstruction in the game "honor of kings". In *2020 3rd International Conference on Humanities Education and Social Sciences (ICHESS 2020)*, pages 131–135. Atlantis Press.

Sharon R Sherman. 1997. Perils of the princess: Gender and genre in video games. *Western folklore*, 56(3/4):243–258.

Alexander Styhre, Björn Remneland-Wikhamn, Anna-Maria Szczepanska, and Jan Ljungberg. 2018. Masculine domination and gender subtexts: The role of female professionals in the renewal of the swedish video game industry. *Culture and Organization*, 24(3):244–261.

Jing Sun. 2020. Gender in chinese video games. *The International Encyclopedia of Gender, Media, and Communication*, pages 1–5.

Jason Wilson. 2022. Honor of kings is getting even bigger with a global release. *Sports Business Journal*.

S Wu, C Yang, and Y Zhang. 2019. The chinese version of moral foundations dictionary: a brief introduction and pilot analysis. *ChinaXiv*, 10(201911.00002).

Yue Xie. 2013. Gender and family in contemporary china. *Population studies center research report*, 13:808.

Siyu Yao and Yumin Chen. 2022. Reconstructing history and culture in game discourse: A linguistic analysis of heroic stories in honor of kings. *Games and Culture*, 17(7-8):977–996.

Meng Zhang. 2012. A chinese beauty story: How college women in china negotiate beauty, body image, and mass media. *Chinese Journal of Communication*, 5(4):437–454.

Suoyi Zhang. 2022. The female body and experience in chinese multiplayer online battle arena games. In *2021 International Conference on Social Development and Media Communication (SDMC 2021)*, pages 1125–1129. Atlantis Press.

Kyrie Zhixuan Zhou, Jiaxun Cao, Xiaowen Yuan, Daniel E Weissglass, Zachary Kilhoffer, Madelyn Rose Sanfilippo, and Xin Tong. 2023a. "i'm not confident in debiasing ai systems since i know too little": Teaching ai creators about gender bias through hands-on tutorials. *arXiv preprint arXiv:2309.08121*.

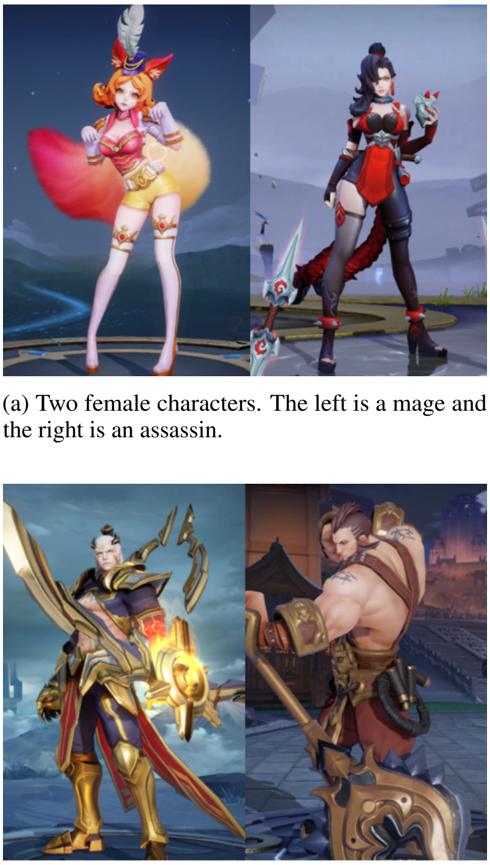
Kyrie Zhixuan Zhou and Madelyn Rose Sanfilippo. 2023. Public perceptions of gender bias in large language models: Cases of chatgpt and ernie. *arXiv preprint arXiv:2309.09120*.

Zhixuan Zhou, Bohui Shen, Franziska Zimmer, Chuanli Xia, and Xin Tong. 2023b. More than a wife and a mom: A study of mom vlogging practices in china. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, pages 56–63.

Zhixuan Zhou, Jiao Sun, Jiaxin Pei, Nanyun Peng, and Jinjun Xiong. 2022a. A moral- and event-centric inspection of gender bias in fairy tales at a large scale. *arXiv preprint arXiv:2211.14358*.

Zhixuan Zhou, Zixin Wang, and Franziska Zimmer. 2022b. Anonymous expression in an online community for women in china. *arXiv preprint arXiv:2206.07923*.

A Example Female and Male Characters



(a) Two female characters. The left is a mage and the right is an assassin.

(b) Two male characters. The left is an archer and the right is a tank/warrior.

Figure 4: Example female and male characters in Honor of Kings for a visual comparison.

The Great Digital Humanities Disconnect: The Failure of DH Publishing

Emily Öhman
Waseda University
ohman@waseda.jp

Michael Piotrowski
University of Lausanne
michael.piotrowski@unil.ch

Mika Hämäläinen
Metropolia University of Applied Sciences
mika.hamalainen@metropolia.fi

Abstract

We discuss the disconnect in interdisciplinary publishing from a disciplinary divide perspective as to how research is expected to be presented and published according to disciplinary conventions. We argue that this divide hinders interdisciplinary collaboration and even more so the dissemination of research results from interdisciplinary projects to other interdisciplinary researchers. The disconnect is not simply theoretical but also encompasses practical considerations such as manuscript creation standards. The disconnect can also be detrimental to academic careers in terms of evaluations by peers on funding and tenure committees as well as peer reviews. With this analysis, we want to foster further discussion about the state of academic publishing from a digital humanities perspective.

1 Introduction

Different academic disciplines have different cultures and traditions, notably different standards and expectations for what constitutes acceptable research within the discipline, but also how this research is to be presented and published. This, in turn, defines what constitutes a good "track record" for researchers, which determines their career prospects (hiring, tenure, and promotion) in this discipline. In fact, the definition and assurance of quality standards could be said to be the main purpose of disciplines, their *raison d'être*. These standards are enforced through institutions that are able to make decisions about funding, hiring, teaching, etc. On this level, all disciplines are effectively in competition, and cultures and traditions thus also have important social roles that may have little to do with the actual quality of research and more with staking out claims in order to influence the allocation of resources (see, e.g., Becher and Trowler, 2001).

Although the cultures and traditions of established disciplines can be arcane and difficult to navigate (absorbing the discipline's culture is an important, though mostly implicit, part of the training in a discipline), the expectations are—in principle—known or at least knowable. However, the picture changes radically when we consider *interdisciplinary* research, and even more so if the work is not just a clearly delimited collaboration of researchers from two or more disciplines, but literally "between" disciplines, i.e., in a kind of academic no man's land that is not ruled by any discipline.

In both cases, researchers are confronted with and need to be aware of the publishing conventions of different disciplines if they want to benefit from their work: they need to ensure that it is visible in the right community, and that is cited or otherwise acknowledged, e.g., for obtaining tenure. The phenomenon is well known as "publish or perish" (Hammarfelt and De Rijcke, 2015).

Interdisciplinary researchers are also confronted with practical issues when considering publication venues. When language technology researchers on Reddit[1] were asked the question: "Would you submit a paper to a venue that did not allow LaTeX/demanded .doc?" specifying the focus as interdisciplinary research (askja, 2023) one scholar commented:

> It depends on the "weight" of the journal. If that's a journal where I'd plan to do some research work explicitly targeting that journal, then requiring .doc would be inconvenient but not that important. But if that is research work which I'd do and then, when it's mostly done, see what's the best place for it – then by the time I'd consider that journal, the content would be already mostly written in LaTeX and then I'd be hesitant to rewrite it in Word.

[1] https://www.reddit.com/

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 132–137
December 1-3, 2023 ©2023 Association for Computational Linguistics

The comment echoes two challenges for the dissemination of interdisciplinary research. The question of where to submit one's work seems obvious and largely unavoidable in interdisciplinary research. The choice of the writing tool, however, seems to be merely a practical question; yet in the case of digital humanities as a field between computing and the humanities, it tends to be a question that goes far beyond personal preferences or technical merits. The question of LaTeX vs. Microsoft Word can be considered a "fault line" in digital humanities, which aligns with numerous other divisions inside the supposed "big tent."

"Big Tent Digital Humanities" was the theme of the DH 2011 conference, and the "big tent" metaphor has been used since to emphasize the diversity, openness, inclusiveness, and fluidity of digital humanities. While well-meaning, already at the time some scholars noted potential problems with this notion. For example, Svensson cautioned that "[e]ven if the big-tent vision of the digital humanities gives the field a sense of openness and invitation, it does not necessarily remove institutional predispositions and thresholds or make the field into a blank slate" (Svensson, 2012, 47); he remarked that, consequently, "there is a risk that a wealth of traditions and perspectives are subsumed and conflated in a tent primarily keyed to one particular tradition" (Svensson, 2012, 45).

This seems in fact to be the case—otherwise, why highlight the fact that the Computational Humanities Research conference,[2] established in 2019 by a group of scholars who do computationally oriented work in DH, has adopted "practices that align with norms in computer science and linguistics (e.g., submission of 6- or 12-page papers, exclusive use of LaTeX)" (Dombrowski, 2023, 138) for their conference?

In this position paper, we will explore the disconnect between these two flavors of DH from a research dissemination perspective. In addition to discussing the difficulties in interdisciplinary research and how the lack of disciplinary coherence reflects on the evaluation of research outputs, we focus on some practical issues within the publication disconnect: what open science means for the different disciplines involved including "publish or perish", the authorship and readership of different publication venues, document processing, and finally what the impact will be for digital humanities as a field going forward.

2 Disciplinary Backgrounds

As most publication venues are associated with a specific discipline, the work to be published will be evaluated according to the standards of this discipline, both with respect to the research *and* the formal requirements concerning its presentation. But as interdisciplinary work is not, or at least not entirely, within a single discipline, this raises significant problems for the evaluation of the work—what constitutes appropriate peer review for interdisciplinary research and who is qualified to judge it? (Bammer, 2016; McLeish and Strang, 2016)—and it has been shown that interdisciplinary research has consistently lower funding success (Bromham et al., 2016). Additionally, the lack of disciplinary conventions means that it can in practice be difficult to find reviewers with appropriate disciplinary knowledge of all the fields involved, which can lead to interdisciplinary research being unfairly evaluated from the perspective of only one of the involved fields, and the difficulty in recruiting appropriate reviewers can often delay the publication process further.

In traditional humanities disciplines, the main research output is journal papers of 15–25 pages (or even longer), edited volumes, and books. It takes a long time to write such manuscripts and it can take years from submission to publication. Typically, there is no document template for submission or publication; instead, there are extensive lists of margin sizes, comma use, capitalization rules, bibliographic guides, and many more submission guidelines. The submission is in the form of a .doc(x) file. Conferences in the humanities are primarily networking events, where scholars present work in progress. These conferences tend to only have abstract submissions (a typical limit is 500 words), which may or may not be published in a book of abstracts.

In language technology, natural language processing (NLP), and computational humanities, the prestige of publication venues is often reversed, with conference proceedings being the main and most prestigious venue for disseminating information; publication in journals is generally considered too slow and low-impact. There are often two categories of conference papers: short papers (4 to 6 pages) and long papers (8 to 12 pages).

[2] https://computational-humanities-research.org/

The papers are indexed and double-blind peer-reviewed. The delay between submission and publication is usually between 3 to 6 months, depending on the specific venue. In these more computational disciplines, LaTeX is overwhelmingly used for submissions—these days, a direct link to a collaborative template on Overleaf[3] is even typically provided. Nevertheless, Word templates are often also available.

Preprints are strongly encouraged and sometimes required in computational fields, further speeding up the information dissemination process. As an example, arxiv, the main preprint service for computer science and other STEM fields received 20,170 submissions in October 2023. Another study found that of arxiv preprint papers submitted in 2017-2018 77% were later published in peer-reviewed venues (Lin et al., 2020). However, preprints in the humanities are mostly frowned upon (Laporte, 2017) and if not outright forbidden it is common that referencing preprints is discouraged in humanities journals with only 45% of humanities journals allowing preprints and nearly all computer science journals allowing them (Klebel et al., 2020).

Digital Scholarship in the Humanities and *Digital Humanities Quarterly* could be considered two of the main journals within DH. However, the articles published in them lean heavily towards digitization (roughly 80%) rather than computational humanities, and much of the actual computational content is limited to stylometrics (Roth, 2019; Piotrowski, 2020). Digitization, especially of cultural heritage, is certainly a part of digital humanities, but the imbalance is striking. Where are the computational humanities papers published, if not in these journals? Are we perhaps after all not all in the same "big tent" despite disciplinary gaps?

3 The Disconnect

We argue that there is a growing theoretical and practical disconnect in DH due to widespread hostility towards certain types of computational approaches and other customs from computational fields. As a result, current DH can to a large extent be characterized as pseudo-interdisciplinary: it is in fact largely dominated by traditional humanities practices, in particular with respect to publishing, which, whether consciously or subconsciously, tend to exclude interdisciplinary researchers from fields such as computer science, computational linguistics, and even computational humanities.

3.1 Open Science

Most academics in any discipline would agree that *Open science* is a good thing allowing everyone access to research results and makes these results more transparent. However, for most humanities scholars open science in practice tends to be limited to paying open-access journal publication fees. Corpora and digitized datasets in the humanities are often locked behind online interfaces or even CD-ROMs, perhaps mostly due to copyright issues, but undeniably also disciplinary culture. On the other hand, in every aspect of computational research, it is highly encouraged for everything to be made publicly available from preprints, to data, and code, and this openness and accessibility is often a part of the peer-review criteria. In combination with the different approaches to the analysis of results, this leads to a situation where the more analysis-focused results rarely face methodological scrutiny from reviewers as it is not expected or possible in most cases due to lack of access to code and data.

3.2 Publish or Perish in a Faux-Interdisciplinary Context

It is unfortunate that many researcher are faced with the choice of "publish or perish," as this leads to insular publication practices. It also means researchers feel compelled to pump out articles at a pace at which it is difficult to maintain academic rigor. "Publish or perish" is an issue in all disciplines; however, the publishing disconnect between digitized and computational humanities (i.e., within the "big tent" of digital humanities) exercebates the problem in DH.

Universities and academic journals both contribute to the pervasive culture of "publish or perish." Facing budgetary pressures, institutions depend on prestige to attract research funding, and one of the easiest ways to increase prestige (as measured by rankings) is to be highly visible in prestigious journals.

It is easy to dismiss "publish or perish" as an old aphorism that academics use to complain about their working conditions, but there is ample evidence (e.g., the replication crisis) that the longer this unhealthy pressure persists, the greater the risk to research integrity. As the researchers start to suffer, and the cracks begin to appear, we can see real consequences: in an attempt to increase publi-

[3] https://www.overleaf.com/

cation metrics, researchers split up project results into “minimum publishable units,” when one paper would have sufficed, join each others’ publications as co-author, publish only research with positive results, or even resort to forgery. Not all of these practices are necessarily bad as such, and splitting up papers into multiple papers can help clarify specific contributions and increase the citability of a paper, however, when viewed as a whole or as a method to “game the system" the ethics of such practices become murky.

3.3 Disciplinary and Interdisciplinary Publication Venues

Most academics are vaguely aware of the fact that publication practices vary by field. However, many are woefully ill-equipped to evaluate the publication record of someone from a different field. Computational sciences expect rapid publication of preprints on the one hand, and on the other end of the spectrum, humanities scholars are expected to take years to write whole books. When these two disciplines come together, which disciplinary background is used to evaluate research output and results?

The choice of publication venue is an important one as it decides not only where your research will be published, but also who will review and who will read about it. This means that journals that call themselves DH but only accept manuscripts structured and created using standards from one end of the interdisciplinary spectrum will be overlooked by those closer to the other end.

3.4 Document Standards as Gatekeeping and Virtue Signaling

The state of the art in scholarly publishing (even when only considering the technical aspects) is appalling. When PDF output is required, LaTeX remains more or less the only comprehensive authoring solution. Writing a paper for, say, an ACM[4] or ACL[5] conference is easy: there are official document templates, you literally just have to write your paper. The point here is not that LaTeX is “better,” but rather that there is a clearly defined path to the submission, and authors do not have to concern themselves with the formatting of the document or the references: this is all taken care of automatically. This also means that no conversion and no manual interventions are required, which tend to introduce errors. In addition, the system is open source and highly portable, you can use it on any platform and with any editor you want.

In the humanities, however, an expensive license for Microsoft Word is usually required and authors have to manually adjust numerous settings and manually ensure that their submission conforms to the guidelines. Thus, although LaTeX is older, it is much better suited for modern publication practices, including automatic compilation of the final product (proceedings, journal). This is not to say that it is perfect, but it is perhaps the best we have at the moment. LaTeX comes with its own host of issues, specifically since the end product, a PDF document, suffers from loss of semantic information that is only available in the source code (Piotrowski, 2016). It has been noted that the requirement to submit a Word document might in some cases simply be due to the editors being unaware of the possibility of LaTeXsubmissions and a request to submit using LaTeXmight be granted especially since all major publishers provide LaTeXtemplates including Springer, Elsevier, Wiley etc. (Jensen, 2018). Nonetheless, none of the purely DH journals, and almost none of the DH conferences offer a LaTeXoption despite this undoubtedly being an issue that has been raised by contributing authors. LaTeXtemplates reduce the workload of both editors and authors so it seems strange to deny contributions written in LaTeX.

XML formats have many potential advantages, but the use of TEI (Text Encoding Initiative) in some DH contexts seems akin to virtue signaling and is far from practical, useful, or even in the spirit of TEI: the conferences that use TEI do *not* actually accept submissions in TEI, but require authors to use the DH Convalidator tool, which converts Word documents to TEI format, so the paper must be written in Word, and it cannot contain more information than available in Word, and it is not even published as a TEI document, but as PDF. Effectively, the semantic information painstakingly extracted from graphically typeset texts (.docx) is eventually completely erased. This marks the other end of the spectrum of the art of digital publishing (Cremer, 2018). The *Other* category in table 1 almost exclusively refers to these types of submissions.

There is a widespread fear of LaTeX in the humanities; this is to some extent understandable,

[4] https://www.acm.org/publications/proceedings-template

[5] https://2023.aclweb.org/calls/style_and_formatting/

Venue	Type	LaTeX	.docx etc	Other	Abstract only
Association for Computational Linguistics (ACL)	Umbrella for NLP conferences	O	O	X	X
Digital Humanities Quarterly	DH journal	X	O	O	N/A
Digital Scholarship in the Humanities	DH journal	X	O	O	N/A
Digital Humanities	DH conference	X	O	X	O
European Association for Digital Humanities	DH conference	X	O	O	O
Digital Humanities in the Nordic and Baltic countries	DH conference	O	O	X	X/O
Computational Humanities Research	CH conference	O	X	X	X
Journal of Data Mining & Digital Humanities	DH/NLP journal	O	O	O	N/A
International Journal of Digital Humanities	DH journal	O	O	X	N/A
Humanist Studies & the Digital Age	DH journal	X	O	X	N/A
Journal of Digital History	DH Journal	X	X	O	N/A

Table 1: ACL includes all ACL-affiliated conferences such as EACL, EMNLP, CoLING, etc. as well as co-located workshops. Abstract-only means that only a book of abstracts will be published. "Other" almost exclusively refers to the use of the DH Convalidator tool, except for DHQ where direct XML/TEI submissions are possible, and the Journal of Digital History which only accepts Python Notebooks.

even though it takes no more than 30 minutes to learn the basics of LaTeX (or a tool like Pandoc[6]), as it is not part of traditional humanities curricula and one study found that just over 20% of humanities scholars were comfortable using TeX or other markup languages including XML and TEI (Bonn and Swatscheno, 2017). In that particular instance, the finding led to a decision to only accept Word documents.

The hostility against LaTeX (or, in fact, anything that is *not* Word) in large parts of DH—a field that commonly describes itself as located at the intersection between computer science and the humanities and that prides itself on its interdisciplinarity, inclusiveness, and progressiveness—is a different story, though. One possible explanation is that it, like the refusal to define DH, serves a gatekeeping function: as Piotrowski (2020) argues, humanities scholars that wield DH as "a term of tactical convenience" (Kirschenbaum, 2014) to obtain a vanguard status in their discipline are natural wary of potential intruders that could strip them of this status.

Figure 1 presents an overview of different digital humanities and computational humanities venues and the file types expected of submissions. All computationally oriented conferences (ACL, ACM, CHR) accept LaTeXsubmissions and provide templates - most of them provide templates for Word as well. On the other hand, very few DH venues accept LaTeX. Notable exceptions include the Journal of Data Mining and Digital Humanities which requires preprint submissions which means any submission type is acceptable - and later provide both LaTeXand Word templates, the Digital Humanities in the Nordic and Baltic Countries conference which accepts both and allows both abstract-only and short and long paper submissions. Noteworthy are also the Digital Humanities Quarterly which does not accept LaTeXbut accepts XML and TEI submissions **not** generated with the DH Convalidator tool, and the Journal of Digital History, which only accepts .ipynb notebooks using their template. Most DH conferences do not publish proceedings beyond a book of abstracts, whereas no computational conference allows abstracts with the exception of lightning talks at the Computational Humanities Research conference.

4 Concluding Remarks

This position paper is a call to action. We believe that there is no future for DH if this disconnect is not addressed. Instead of waiting for the "big tent" to collapse—to the detriment of everybody who is in it—we should work to establish *computational humanities* as a discipline in its own right, that sets its own standards and evaluation criteria. In the end, this will be the only way to ensure adequate recognition of computational research in the humanities. This does not have to mean the collapse of digital humanities, but instead a strengthening of the position of digital humanities as a field separate from traditional humanities with appropriately adjusted evaluation criteria and mutually agreed upon publication practices that are neither cumbersome nor slow to use or publish.

References

askja. 2023. Reddit post.

Gabriele Bammer. 2016. What constitutes appropriate peer review for interdisciplinary research? *Palgrave Communications*, 2(1).

[6] https://pandoc.org/

Tony Becher and Paul R. Trowler. 2001. *Academic Tribes and Territories: Intellectual Enquiry and the Cultures of Discipline*, 2^{nd} edition. Open University Press, Buckingham, UK.

Maria Bonn and Janet Swatscheno. 2017. Humanities without walls: Understanding the needs of scholars in the contemporary publishing environment.

Lindell Bromham, Russell Dinnage, and Xia Hua. 2016. Interdisciplinary research has consistently lower funding success. *Nature*, 534(7609):684–687.

Fabian Cremer. 2018. Nun sag, wie hältst Du es mit dem Digitalen Publizieren, Digital Humanities?

Quinn Dombrowski. 2023. Does coding matter for doing digital humanities? In James O'Sullivan, editor, *The Bloomsbury Handbook to the Digital Humanities*, chapter 13, pages 137–145. Bloomsbury, London.

Björn Hammarfelt and Sarah De Rijcke. 2015. Accountability in context: Effects of research evaluation systems on publication practices, disciplinary norms, and individual working routines in the Faculty of Arts at Uppsala University. *Research Evaluation*, 24(1):63–77.

Lars Christian Jensen. 2018. Guest blog post: Latex for the humanities.

Matthew G. Kirschenbaum. 2014. What is "Digital Humanities," and why are they saying such terrible things about it? *differences*, 25(1):46–63.

Thomas Klebel, Stefan Reichmann, Jessica Polka, Gary McDowell, Naomi Penfold, Samantha Hindle, and Tony Ross-Hellauer. 2020. Peer review and preprint policies are unclear at most major journals. *PLoS One*, 15(10):e0239518.

Steven Laporte. 2017. Preprint for the humanities–fiction or a real possibility? *Studia Historiae Scientiarum*, 16.

Jialiang Lin, Yao Yu, Yu Zhou, Zhiyang Zhou, and Xiaodong Shi. 2020. How many preprints have actually been printed and why: a case study of computer science preprints on arxiv. *Scientometrics*, 124(1):555–574.

Tom McLeish and Veronica Strang. 2016. Evaluating interdisciplinary research: the elephant in the peer-reviewers' room. *Palgrave Communications*, 2(1):1–8.

Michael Piotrowski. 2016. Future publishing formats. In *Proceedings of the 2016 ACM Symposium on Document Engineering*, DocEng '16, page 7–8, New York, NY, USA. Association for Computing Machinery.

Michael Piotrowski. 2020. Ain't no way around it: why we need to be clear about what we mean by "digital humanities". SocArXiv.

Camille Roth. 2019. Digital, digitized, and numerical humanities. *Digital Scholarship in the Humanities*, 34(3):616–632.

Patrik Svensson. 2012. Beyond the big tent. In Matthew K. Gold, editor, *Debates in the Digital Humanities*, pages 36–72. University of Minnesota Press.

Explorative study on verbalizing students’ skills with NLP/AI-tool in Digital Living Lab at Laurea UAS, Finland

Asko Mononen
Laurea University of Applied Sciences, Finland
asko.mononen@laurea.fi

Abstract

This explorative study tested Laurea UAS students’ (N=16) abilities to verbalize their skills, before and after the study unit “Digital Analytics and Consumer Insights”. Before the study unit the students listed their skills unaided and afterwards with help of Careerbot AI -service. The findings indicate that the intervention increased both quantity and quality of the skills verbalized, relevant to the learning objectives and generic, 21st century skills.

1 Introduction

The purpose of this explorative study was to research if the students can verbalize their skills and competences better with the help of Careerbot AI -service than without it.

Laurea University of Applied Sciences in Helsinki region in Finland has a learning environment called Digital Living Lab (DLL), focusing on real-life project-based studies with partner organisations. The DLL aims to support the acquisition of “21st century skills”, working life skills focusing on digital service development.

Trilling et al. (2009) defines 7C's of 21st century skills as:

- Critical thinking and problem solving
- Creativity and innovation
- Collaboration, teamwork, and leadership
- Cross-cultural understanding
- Communications, information, and media literacy
- Computing and ICT (information and communication technology) literacy
- Career and learning self-reliance

In August 2023, five day design sprint for study unit “Digital Analytics and Consumer insights” (DACI) was executed in English. This study unit is part of elective studies. The author was the responsible lecturer and the head facilitator during the whole hybrid event. Six other facilitators and subject-matter experts supported partially.

Learning objectives for this 5-credit point (ects) study unit were the following: “After the study unit, the student is able to:

- recognize the consumer behaviour offline and online (per main demographics)
- plan data collection points and methods online
- analyse data (e.g., aggregation, trends, comparison)
- visualize results (e.g., dashboards)
- plan consumer activation methods based on data” (Laurea, 2023)

2 Sample

The participating students (N=16) consisted of 15 bachelor and 1 master-level students.

9 of them participated face-to face, and 7 online.

Study fields were Business Management (n=11), Business Information Technology (n=3), Hospitality management (n=1) and Service Design (n=1).

The age groups were in the following categories: 25 or less (n=6), 25-34 (n=2), 35-44 (n=7) and 45-54 (n=1).

Their previous degree was vocational level (n=1), high school/matriculation exam (n=7),

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 138–142
December 1-3, 2023 ©2023 Association for Computational Linguistics

bachelor level (n=7) and master's level of more (n=1).

The participants' native language was mixed, Finnish (n=8) and non-Finnish (n=8, several languages, not specified here for privacy reasons).

The participants' average work experience from knowledge intensive work was 6,36 years. Knowledge intensive work here was defined here as "creative work, requiring complex thinking and communication, vs. routine or manual work".

3 Related work

The use of artificial intelligence (AI) tools in helping higher education students to verbalize their skills and competences in a job market language (as defined in job ads) has not been researched much yet.

Mononen et al. (2023) conceptualized "forecasted self", future-oriented digital twin, where a student can explore several future selves equipped with new, acquired skills for projected future jobs with Careerbot AI -service.

Westman S., & Mononen A., et al. (2021) discussed the prospects for career coaching, with four AI maturity levels in career guidance: 1. AI-aware guidance, 2. AI-informed guidance, 3. AI-integrated guidance and 4. AI-transformed guidance.

Transversal and transferable skills and competences were defined in Transval-EU project (2021) as generic working life skills, soft skills, and employability skills.

Brown and Souto-Otero (2020) analyzed 21 million job ads in the UK and found that employers are most likely to focus applicants "job readiness", demonstrating both generic, soft skills and technical requirements.

Brown and Hesket (2004) talked about potential job candidates' fit with organization through "narrative of employability" boosted by non-work-related skills acquirement ct. qualifications.

Claro et al., (2012) defined ICT literacy above the mastery of ICT applications, to include "higher-order thinking processes", like problem solving of information, communication, and knowledge tasks in an ICT context, relevant to learning context in the knowledge society.

21st century skills were popularized by Trilling & Fadel (2009), rooting back to 1980's.

4 Technology and data

The tool used for students on verbalizing their skills is called "Careerbot". This webservice interface has been developed "for helping 34 000 students to pursue their dream careers with the help of AI" by 3AMK. 3AMK is a strategic alliance of Laurea, Haaga-Helia, and Metropolia universities of applied sciences in Helsinki region, Finland. (3AMK.fi, 2023)

The AI behind this webservice is called "Graphmind" and built by Finnish tech company HeadAI Ltd. Graphmind is a Graph Machine Learning -based semantic computing framework accessible via REST-API for Careerbot -service. (Mononen et al., 2023).

In the first phase Graphmind has been taught with unstructured data (millions of news) and e.g., European Skills, Competences, Qualifications and Occupations (ESCO) classification, and in the second phase with reinforced learning. (Headai Ltd., 2023)

The main data source for Careerbot is job market data in Finland (Työmarkkinatori, MOL and Duunitori/employment services) with over 400 000 job ads on a yearly basis since January 2018.

The other data sources are 3AMK course data, Theseus -theses data from Finland and global directory of open access journals (DOAJ) but they were not used for this study.

In the Careerbot AI -service the students can create their skills profile, a digital twin for skills, "forecasted self". The skills are defined pragmatically, as the words are stated in the job ads.

The user experience flow in Careerbot AI-service works in the following way: a) login to the system, b) create a new skills profile, c) start typing in the skills words individually (and get suggested related skills words), optionally d) copy-paste personal cv to text field from which the skills words are retrieved, e) select relevant skills word from suggested skills word list, f) look for jobs in Finland based on area, time and your skills profile (selected skills words), and g) further update your skills profile from selected job (Finnish job ads "soft- and hard skills" that were still missing from one's skills profile).

There were some bugs encountered during the session with students. However, every student managed to perform the given assignment on verbalizing their own skills with the help of Careerbot AI -service.

5 Methods

During the sprint week, the students were assigned into five teams of 3-4 people randomly, 2 teams online and 3 face-to-face in the Digital Living Lab. Teams could choose their team assignments freely, along the learning objectives.

Strictly speaking this was a quasi-experiment, since 3 students were allowed to change their assigned study method (face-to-face or online) to another one for personal reasons, after ethical consideration.

This sub-study focused on students' ability for skills verbalization, not on the skills development per se. The other sub-study will be reporting the findings for perceptions on the course.

The briefing for the students is stated below:

> "...Please answer honestly to this questionnaire, how do you feel about the claims right now. These are your personal views, there are no right or wrong answers, and these answers will not affect your study unit grading…"
>
> "YOUR DACI -RELATED SKILLS
> List down YOUR CURRENT SKILLS after the "digital analytics & consumer insights" -study unit. You can list as many as you can. Please list skills one per line in the open text field below.
> (Addition in POST-questionnaire:) List down spontaneously and use also the 3AMK.FI/CAREERBOT AI-service for listing your skills.
> (Open ended text -field for answers)"

In the PRE-questionnaire, the students were given only written instructions via email and online questionnaire with open text field for answer.

Before the POST-questionnaire, on the last day of study unit, the students were introduced to usage of Careerbot AI -service for 10 minutes. The students were instructed on how to create a new skills profile, look for jobs based on it, and further educate their skills profiles based on missing skills found on the job ads. The students used 30 minutes for the assignment and filled in their skills words to POST-questionnaire with open text field.

6 Results

The results of verbalizing student's skills with the PRE vs. POST experiment setup indicate a clear increase of quantity and quality of the skills words for every participated student (N=16).

	PRE	POST	DIFF.
Average	3.56	13.94	10.38
Median	3.5	12.5	9
Std Dev	2.60	6.06	3.46

Table 1: DACI-study, quantity of skills

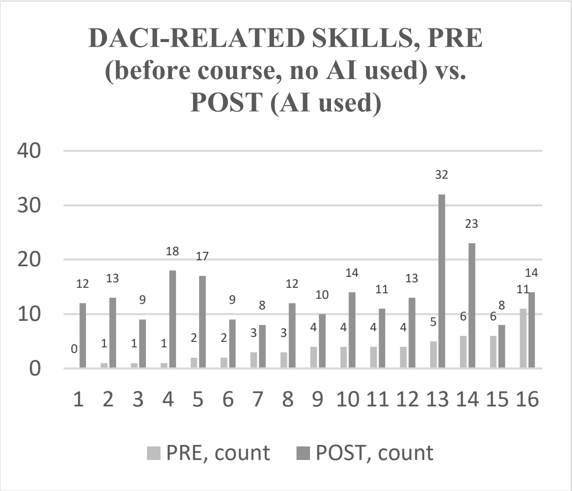


Figure 3: DACI-related skills count, PRE vs. POST (with AI used) per student 1-16.

The quality of the verbalized skills was evaluated based on the relevance to study unit learning objectives and 21st century skills by the author. From the students open text field -answers the skills words were manually extracted and copied as a list to https://www.wordclouds.com/ from which the frequencies were extracted, synonyms combined and visualized.

Before the study unit the students listed their skills and their frequency as following (top 20):

#	skill	#	skill
6	research*	2	problem*
4	thinking*	2	market*
4	design*	2	making*
3	excel*	2	journey*
3	data*	2	customer*
3	consumer*	2	critical*
2	solving*	2	communication*
2	skills*	2	behavior*
2	segmentation*	2	analysis*
2	problem-solving*	1	visualisation*

Table 2: PRE-questionnaire skills list & freq.
*can be any character/word as continuation.

After the study unit, with the help of Careerbot AI - service, the students listed their skills and their frequency as following (top 20):

#	skill	#	skill
42	Data*	8	Research*
22	Marketing*	7	Communication*
19	Analysis*	7	Insight*
15	Customer*	7	Management*
13	Consumer*	7	Service*
11	Analytics*	7	Trends*
9	Business*	7	Visualization*
9	Digital*	6	Development*
8	Collection*	6	Product*
8	Excel*	5	Design*

Table 3: POST-questionnaire skills list & freq.
*can be any character/word as continuation.

7 Discussion

The purpose of this exploration was to research if the students can verbalize their skills better with help of Careerbot AI -service than without it.

As a conclusion, the quantity verbalized skills increased for all the participated students, on average from 3,56 to 13,94 (median from 3,5 to 12,5). The students learned to elaborate more on the topic of the study unit.

The quality of the verbalized skills also increased. The mentioned skills can be seen relevant from employment point of view, since they all appear in job ads (=data source). Also, more students mentioned keywords relevant to the learning objectives and 21st century skills (generic, soft skills) after the sprint week.

Evaluating the extent to which the difference was due to teaching method vs. AI is not clear. The earlier sessions using Careerbot AI-service with the students have produced results of similar direction, but they have not been documented systematically.

Both face-to-face and online groups received the same introduction and guidance for the AI service use simultaneously. Face-to-face group improved slightly more during the week (3,89 →12,3 vs. online 3,14→15,2 skills). Sample size is too small for conclusions on the difference.

Based on the earlier feedback of industry partners of Laurea UAS in autumn 2022, the graduates have had sometimes difficulties in verbalizing and therefore “selling” their skills to the recruiters. For this, the use of AI in addition to teaching and coaching can be useful. This is in line with Brown, et al. (2004) who highlighted “narrative of employability” for job candidates.

This experiment was using AI as a personal tool, reaching at most the level 2/4 “AI-informed guidance” in maturity model, Westman S., et al. (2021). So, there is still a gap for reaching the higher levels in AI maturity in coaching.

However, to confirm these initial findings, more studies are needed with larger sample sizes. Also, more studies are needed with other tools in comparison, including traditional human guided career coaching and other tools, like latest versions of generative AI (ChatGPT, etc.).

Acknowledgments

Thank you for Laurea UAS for the research permission granted on June 15, 2023.

References

Brown, P., Hesketh A. (2004). The mismanagement of talent: Employability and jobs in the knowledge economy. Oxford University Press.

Brown, P., Souto-Otero, M. (2020). The end of the credential society? An analysis of the relationship between education and the labour market using big data. Journal of Education Policy 35 (1), pp. 95-118. https://doi.org/10.1080/02680939.2018.1549752

Claro, M., Preiss, D. D., San Martín, E., Jara, I., Hinostroza, J. E., Valenzuela, S., Cortes, F., & Nussbaum, M. (2012). Assessment of 21st century ICT skills in Chile: Test design and results from high school level students. *Computers & Education*, *59*(3), 1042-1053. https://doi.org/10.1016/j.compedu.2012.04.004

European State of the Art Report, VALIDATION OF TRANSVERSAL SKILLS ACROSS EUROPE (December, 2021). https://www.transvalproject.eu/wp-content/uploads/2022/03/D2.1_State-of-the-Art-Report_EN_public.pdf

Headai Ltd. *No magic, just science.* Retrieved October 2, 2023 from https://headai.com/science/

Laurea University of Applied Sciences. *Digital Analytics and Consumer Insights, study guide*. (October 2, 2023). https://ops.laurea.fi/212701/fi/68154/209690/2692/0/27172

Mononen, A., Alamäki, A., Kauttonen, J., Klemetti, A., Passi-Rauste, A., & Ketamo, H. (2023). *Forecasted Self: AI-Based Careerbot-Service Helping Students with Job Market Dynamics.* The 9th International Conference on Time Series and Forecasting, 99. https://doi.org/10.3390/engproc2023039099

3AMK, homepage. Retrieved November 11, 2023 from https://www.3amk.fi/en/welcome-to-3amk-fi/

Trilling, B., & Fadel, C. (2009). *21st Century Skills: Learning for Life in Our Times.* Wiley.

Westman, S., Kauttonen, J., Klemetti, A., Korhonen, N., Manninen, M., Mononen, A., Niittymäki, S., & Paananen, H. (2021). *Artificial Intelligence for Career Guidance – Current Requirements and Prospects for the Future,* 9(4). https://doi.org/10.22492/ije.9.4.03

Combating Hallucination and Misinformation: Factual Information Generation with Tokenized Generative Transformer

Sourav Das, Sanjay Chatterji, and **Imon Mukherjee**
Department of Computer Science and Engineering
Indian Institution of Information Technology Kalyani
Kalyani, West Bengal, India
{sourav_phd21, sanjayc, imon}@iiitkalyani.ac.in

Abstract

Large language models have gained a meteoric rise recently. With the prominence of LLMs, hallucination and misinformation generation have become a severity too. To combat this issue, we propose a contextual topic modeling approach called Co-LDA for generative transformer. It is based on Latent Dirichlet Allocation and is designed for accurate sentence-level information generation. This method extracts cohesive topics from COVID-19 research literature, grouping them into relevant categories. These contextually rich topic words serve as masked tokens in our proposed Tokenized Generative Transformer, a modified Generative Pre-Trained Transformer for generating accurate information in any designated topics. Our approach addresses micro hallucination and incorrect information issues in experimentation with the LLMs. We also introduce a Perplexity-Similarity Score system to measure semantic similarity between generated and original documents, offering accuracy and authenticity for generated texts. Evaluation of benchmark datasets, including question answering, language understanding, and language similarity demonstrates the effectiveness of our text generation method, surpassing some state-of-the-art transformer models.

1 Introduction

Large language models have become paradigm-shifting research in natural language processing, with their outstanding abilities already demonstrated in a multitude of tasks (Zhao et al., 2023). While the concept of LLMs is not entirely new (Brants et al., 2007), they have obtained computational and performative success in the last couple of years due to the enormous growth of required hardware resources. Despite the research success, an issue with the existing LLMs is that, while continuing a conversation with any LLM-based conversational agent, the rapid shifting of topics and contexts as input prompts often lead the conversation into a logical void. In such a scenario, often some LLM-based agents (for instance, *ChatGPT*[1]) cannot grasp changing context and changing the topic to persuade for delivering an expected response. Others (for instance, *Google Bard*[2]) show multiple drafts of the response, allowing the users to choose the better response as per their liking. One or more such responses can even produce *misinformation*. Another problem being discussed quite a lot about LLMs is the *hallucination* problem (Ji et al., 2023). Here, the received answer is not structured with desirable logic and reasoning, and the flow of the specific content generation can heavily deviate. In addition, some level of intellectual knowledge of semantics and queries is required to explicitly fine-tune the conversation initiation questions in certain ways to get the desired answer. This knowledge is known as prompt engineering.

In this paper, we develop an incremental-learning-based contextual topic modeling algorithm, for the generation of contingently relevant information from a large corpus of research papers related to COVID-19. The use case of our approach is made in the ground truth of the scientific literature on COVID-19, where effective information generation and analysis are vital for understanding and communicating critical insights. We propose *Contextual LDA* (Co-LDA), to extract contextual topics from the training set. We generate four sets of contextually rich topics (*T1* to *T4*) with ten topic words each using the Context Scores derived from the Co-LDA algorithm, categorized into distinctive categories such as *Medical*, *Social*, *Research*, and *Generic* topics for distinct grouping.

Upon selecting the best iteration with the highest context score, we track the original sentences from the training set containing topic words from the resultant topic set. Then a benchmark dataset is ac-

[1] https://openai.com/blog/chatgpt

[2] https://bard.google.com/

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 143–152
December 1-3, 2023 ©2023 Association for Computational Linguistics

quired as a test set and the same steps are performed as with the training set. To generate informative sentences, we construct a tokenizer-based transformer on the existing GPT-3 and call this model the *Tokenized Generative Transformer* (TGT). The extracted topic words are then converted to corresponding masks using the Context Scores before feeding them as inputs in the TGT model. The masked topic words are then converted into a numerical encoded sequence of tokens suitable for model input. This process enables our model to understand and generate contextual fact-based text with better token-level accuracy and semantics.

We evaluate the performance of generated sentences based on the accuracy in terms of contextuality and semantics. Here, we propose the *Perplexity-Similarity Scores* to calculate the pairwise similarity scores between the comparing document sets of original sentences and generated sentences. To improve the accuracy of our computations, variable-length functions are considered for the comparable sentences. A higher matching score indicates better semantic similarity, with more accurate information augmentation. Our experimental framework outperforms most of the compared baseline state-of-the-art LLMs in factual information generation tasks on the same test data.

In this work, we make several fundamental contributions, outlined as:

- We propose **Contextual LDA (Co-LDA)**, a topic modeling algorithm based on incremental learning from data and addressing limitations of the traditional LDA by emphasizing context for more meaningful topic representations.
- We introduce **Tokenizer-based Generative Transformer (TGT)** for information sentence generation, leveraging GPT-3 to overcome hallucination without extensive prompt engineering.
- We also introduce **Perplexity-Similarity Scores** for evaluating accuracy and similarity between original and generated information using variable pairwise distance computation.
- The benchmarking procedure includes evaluation and analysis of our proposed system, based on multiple standard metrics on the public benchmark corpus for the performance demonstration with the comparable SoTA language models.

2 Motivation

Exhaustive research on LLMS has exposed challenges in mitigating hallucination (Ye et al., 2023) and misinformation generation (Pan et al., 2023). Existing LLMs exhibit difficulties in maintaining logical coherence during contextual shifts, leading to misinformation (Karpinska and Iyyer, 2023).

Our motivation lies in addressing these challenges by proposing a method tailored for accurate information generation, particularly in the context of COVID-19 research literature at present. We recognize the limitations of LLMs, notably the hallucination problem (Ji et al., 2023), prompting the development of a solution that overcomes these issues.

To achieve this, we introduce Co-LDA, a contextual topic modeling algorithm. Co-LDA enhances the accuracy of topic representation by considering context, thereby improving the relevance of generated information. By leveraging Co-LDA in conjunction with our Tokenized Generative Transformer (TGT), based on GPT-3, we alleviate the need for extensive prompt engineering, addressing the challenges posed by contextually shifting topics.

The chosen methodology ensures that our model comprehensively captures the nuances of context, leading to more coherent and accurate information generation. Through the integration of Co-LDA and TGT, we aim to provide a robust solution to the challenges associated with hallucination and misinformation in the ground truth of the COVID-19-based information generation.

3 Related Works

The recent flow of NLP research has produced high-standard works for multitask applications. In this section, we follow three major paradigms that closely line up with our work.

3.1 Topic Modeling

The methodologies for topic modeling are similar to dimensionality reduction techniques used for mathematical information. It is often seen as a way of extracting the desired part from the vocabulary. The method of recognizing, modeling, and extracting the topics from any corpus can also result in the thematic representation of the data.

Latent Dirichlet Allocation (2001) is one of the most profound methods for statistically extracting topics from texts (Blei, 2001). It is a measurable graphical model used to establish connections between different reports in the corpus. It is built using *Variational Exception Maximization* (VEM) computations to obtain the most extreme probability measures from the entire text corpus. This can usually be resolved by carefully choosing the most important words. This model follows the idea that each dataset can be represented by a probabilistic propagation of topics and each point can be represented by a probabilistic propagation of words.

Short text topic modeling is the identification of underlying topics within a collection of short text documents. One of the biggest challenges when modeling short text topics is data sparsity. Fewer words appear in the data at any one time, making it difficult to learn the relationships between words and topics. Some researchers introduced the *Topic-Semantic Contrastive Topic Model* (TSCTM), a new framework for modeling short text topics (Wu et al., 2022). TSCTM uses a contrastive learning method to learn relationships between words and topics. This contrasting learning method refines word and topic representations, strengthens the learning signal, and alleviates data sparsity issues.

3.2 Language Generation

The *Generative Pre-trained Transformers* (GPT) (2018) (Radford et al., 2018) has made a breakthrough in NLP. GPT models[3] have achieved remarkable success in various NLP tasks. GPT is from the family of the transformer networks (Vaswani et al., 2017), and comes under the group of Large language models (Zhao et al., 2023). The concept of pre-training a transformer on a large text corpus and fine-tuning the same for a specific task led to the development of these models. The core idea behind the GPT model is to use unsupervised learning to provide a foundation for language understanding, which can then be adapted to various streamlined tasks. Current research potentials in GPT involve addressing biases, improving fine-tuning techniques, and adjusting goals before training.

Large language models are effective for a variety of tasks, recognizing their tendency to generate false information. some researchers focused on assessing LLM's preference for fact-consistent content and introducing a *Factual Inconsistency Benchmark* (FIB) in the context of summarization (Tam et al., 2023). FIB compares LLM results to compare summaries of news articles that are factually consistent with summaries of news articles that are inconsistent. They used human-generated reference summaries that are reviewed for factual consistency and annotated summaries produced by summarization models that are known to be factually inconsistent. The model's accuracy in assessing factual consistency is measured by the proportion of documents that are assigned a high score for a consistent summary of facts.

3.3 Semantic Similarity

Reimers and Gurevych (2019) presented the *Sentence-BERT*, a method for generating sentence embeddings using Siamese BERT networks (Reimers and Gurevych, 2019). The authors used a pre-trained BERT model to encode sentences and learn sentence representations. Using a Siamese network architecture, the model captured semantic similarities between pairs of sentences. Their work contributed to the advancement of sentence embedding techniques for similarity learning. Using a Siamese network architecture, the model with pre-trained BERT learned how to assign sentence pairs to a similarity space in which similar sentences are grouped.

In NLP tasks such as semantic similarity assessment, paired texts often overlap and share components, making accurate semantic assessment difficult. Traditional semantic metrics based on word representations can be confounded by such overlap. To alleviate this problem, Peng et al. (2023) introduced a mask and prediction approach (Peng et al., 2023). Identify words in the *Longest Common Sequence* (LCS) as neighborhood words and predict their position distribution using *Masked Language Modeling* (MLM) from a pre-trained language model. *Neighboring Distribution Divergence* (NDD) metrics quantify semantic distance by measuring the divergence between distributions within overlapping segments.

4 Methodology

We first construct a corpus from a large collection of COVID-19 research literature and perform incremental learning to modify learning features after each topic modeling cycle. We then propose Co-LDA, a contextual topic modeling technique, to extract semantically meaningful topics catego-

[3] https://platform.openai.com/docs/guides/gpt

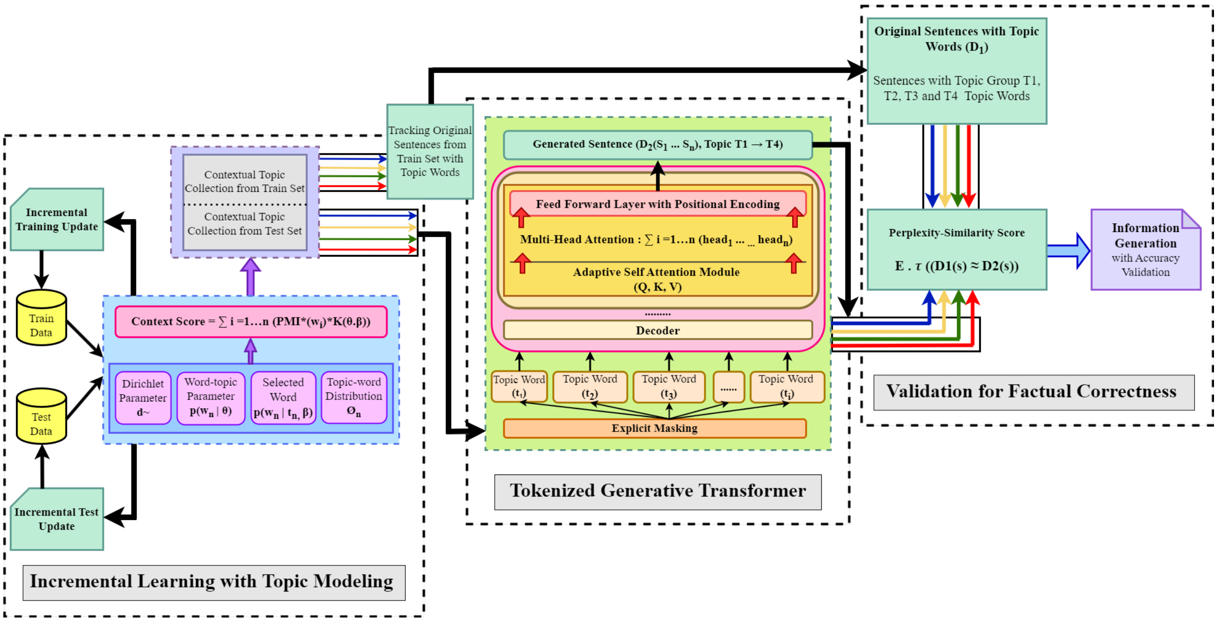

Incremental Training Update
Train Data
Test Data
Incremental Test Update
Contextual Topic Collection from Train Set
Contextual Topic Collection from Test Set
Context Score = ∑ i =1…n (PMI*(w_i)*K(θ.β))
Dirichlet Parameter d~
Word-topic Parameter p(w_n | θ)
Selected Word p(w_n | t_n, β)
Topic-word Distribution Ø_n
Incremental Learning with Topic Modeling
Tracking Original Sentences from Train Set with Topic Words
Generated Sentence (D_2(S_1 ... S_n), Topic T1 → T4)
Feed Forward Layer with Positional Encoding
Multi-Head Attention : ∑ i =1…n (head_1 ... ... head_n)
Adaptive Self Attention Module (Q, K, V)
....
Decoder
Topic Word (t_1)
Topic Word (t_2)
Topic Word (t_3)
......
Topic Word (t_i)
Explicit Masking
Tokenized Generative Transformer
Original Sentences with Topic Words (D_1)
Sentences with Topic Group T1, T2, T3 and T4 Topic Words
Perplexity-Similarity Score
E . τ ((D1(s) ≈ D2(s))
Information Generation with Accuracy Validation
Validation for Factual Correctness

Figure 1: Overview of the proposed system framework. Each phase of the experiment is represented by separating dotted lines. Phases are segregated in order of topic modeling, information generation, and information validation. In phases 1 and 3, the respective topic streams are denoted in identifying shades as Topic T1, Topic T2, Topic T3, and Topic T4.

rized into four groups. These contextual topics are utilized to mask the Tokenized Generative Transformer (TGT), built over GPT-3, to generate factually consistent sentences. To evaluate, we introduce Perplexity-Similarity Scores to measure semantic similarity between the original and generated sentences. Multiple benchmarks demonstrate the effectiveness of our approach over comparable SoTA models.

In this section, we explain the end-to-end experiment pipeline for our work. The overview of the schematic architecture is shown in Figure 1.

4.1 Training Dataset

To construct the training dataset, we gather 5000 research papers denoted as $\mathcal{P} = \{P_1, P_2, \ldots, P_{5000}\}$ from the *arXiv* repository[4], focusing on COVID-19 research published between March 2020 to July 2023. We employ a publication retrieval framework that leverages the arXiv API to conduct a targeted search based on the user-specified topic, denoted as Topic. The topic is interactively provided through the input prompt. The search process is governed by a set of parameters represented as $\mathcal{S} = \{\text{Topic}, \text{MaxResults}, \text{SortCriterion}, \text{SortOrder}\}$, where:

- Topic is the user-specified topic (e.g., COVID-19).
- MaxResults is the maximum number of results to be retrieved (500).
- SortCriterion is the sorting criterion for the search results (e.g., submission date).
- SortOrder is the order in which the results are sorted (e.g., descending with the availability year on arXiv).

The search query returns a set of results denoted as $\mathcal{R} = \{R_1, R_2, \ldots, R_{500}\}$, where each R_i contains metadata and information about a paper, including:

- Title(T_i): The title of the paper P_i.
- Date(D_i): The publication date of paper P_i.
- ID(ID_i): The unique entry ID assigned to paper P_i by the arXiv repository.
- Summary(S_i): A concise summary of the content of the paper P_i, providing insights into its research focus.
- URL(U_i): The URL linking to the original version of the paper P_i, facilitating access to the full text for further examination.

We process these results and organize them in a structured format for subsequent analysis and exploration.

To manage the data, we initialize an empty list $\mathcal{L}$ to serve as temporary containers for each paper. We then iterate through the search results $\mathcal{R}$ and for each paper R_i, the temporary container $\mathit{Container}(R_i)$ is populated with the respective metadata.

[4] https://arxiv.org/

Subsequently, each Container(R_i) is appended to a new, initially empty list denoted as $\mathcal{L}$. Once all the search results have been processed, a data frame is created with the collected data, using column names corresponding to the extracted attributes. The resulting data frame is denoted as *DataFrame*$(\mathcal{L})$.

As new information is continuously surfacing in terms of experiments and results from scientific literature, we utilize additional incremental units for the dataset for continuous learning. When new information is fetched during the API call, the model can be adjusted to learn from the extracted topics without having to completely retrain it from scratch. we assume to represent DataFrame$(\mathcal{L})$ as a collection of data points:

$$\text{DataFrame}(\mathcal{L}) = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\} \tag{1}$$

Here, x_i represents the train (and test) data and y_i represents the previously discussed parameter labels associated with the data. Our goal in incremental learning is to update the parameters Φ using a new subset of data frame $\mathcal{L}_{new}$ while retaining the knowledge learned from previous API calls.

Our key challenge here is to adapt the model's parameters to the new data without forgetting the knowledge gained from the old data. We aim to minimize a loss function *J* that measures the difference between the model's predictions and the baseline parameters. In an incremental setting, we set two components for the loss:

Loss determined on the old data:

$$J_{\text{old}}(\Phi) = \sum_{i=1}^{N_{\text{old}}} \ell(f(x_i; \Phi), y_i) \tag{2}$$

Loss determined on the new data:

$$J_{\text{new}}(\Phi) = \sum_{i=1}^{N_{\text{new}}} \ell(f(x_i; \Phi), y_i) \tag{3}$$

Here, N_{old} and N_{new} represent the number of data points in the old and new datasets, respectively. The function $f(x_i; \Phi)$ represents the model's prediction for input x_i using parameters Φ, and ℓ is the loss function that quantifies the error between the model's prediction and the ground truth.

The overall objective in *incremental learning* is to find a set of updated parameters Φ^* that minimize the combined loss:

$$\Phi^* = \arg\min_{\Phi} \left(\alpha J_{\text{old}}(\Phi) + (1 - \alpha) J_{\text{new}}(\Phi)\right) \tag{4}$$

Here, α is a hyperparameter that controls the balance between preserving knowledge from the old contextual topics (J_{old}) and adapting to the new contextual topics (J_{new}).

4.2 Topic Modeling with Contextual LDA (Co-LDA)

For the improved topic modeling with context, we propose the Latent Dirichlet allocation embedded with *Context Scores* for emphasizing contextuality in extracting meaningful topics from the developed corpus. We call this scheme the Contextual LDA, or Co-LDA. Four Topic Domains or Groups[5]. are observed and derived from this method, corresponding to *T1: Medical Topic*, *T2: Social Topic*, *T3: Research Topic*, and *T4: Generic Topic*. The labeling helps us to compute context scores for different domains. We create an iterative approach to train the Co-LDA model and evaluate its context for retrieving better topic words.

Let us say the datasets containing the research paper summaries are *D*. The dataset is preprocessed to extract the text data by removing stop words, special characters, equations, and diagrams. The preprocessed dataset is denoted by *K* and each of the preprocessed summaries is denoted by k_i. We limit the number of words to k_i a maximum of 500.

$$K_{ij} = \mathit{count}(j \; in \; k_i) \tag{5}$$

where K_{ij} represents the matrix containing $i \times j$ vector representations of the topic words from the training set.

The context score for the topic words is a measure of how well the words from each topic group or domain are related to each other, as well as to the theme of the summary. The context score is typically calculated using a measure of semantic similarity between words. We utilize the *Pointwise Mutual Information* (PMI) (Bouma, 2009) metric to perform this. Higher PMI scores indicate more contextual and meaningful topics. The context score for a set of topics is computed from the PMI score as follows:

$$\mathit{Context\ Score} = \sum_{i=1 \ldots n}^{N} (PMI(w_i) * K(\theta.\beta)) \tag{6}$$

PMI(w_i) represents the PMI score for the topic word w_i. Furthermore, θ is the probability of defining a topic belonging to a summary, and β is the

[5] We may interchangeably use the terms *topic domains* and *topic groups* throughout the paper.

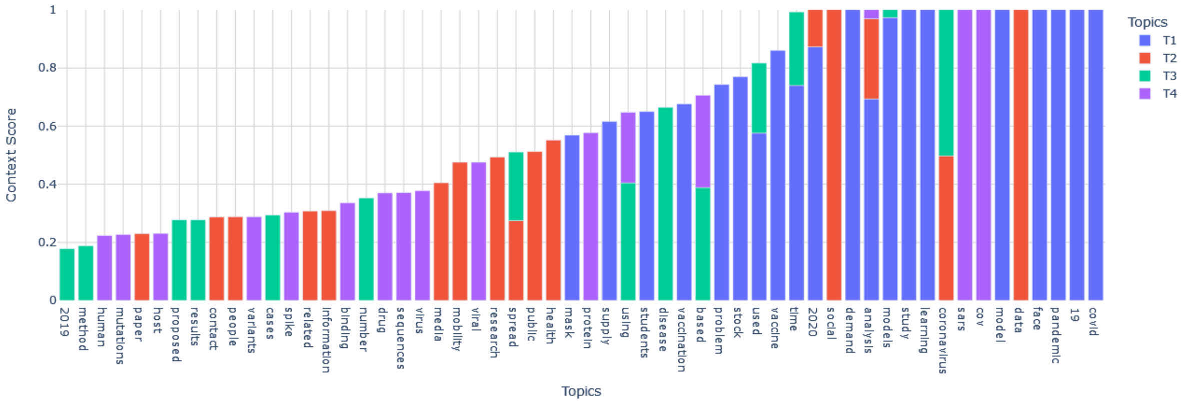


Figure 2: Topic words shown in linear incremental order in terms of Context Scores. The overlapped colors in particular bars represent corresponding topic words that are present in multiple topic groups. It ensures the context sensitivity of the proposed contextual topic modeling.

topic word distribution value. Figure 2 shows a linear increment of PMI scores for a set of topic words from all groups. We record the top ten topic words from each topic group according to the Context Scores for further analysis. The four topics with corresponding topic words are shown in Table 1.

Topic Groups	Topic Words
Topic 1: Medical	pandemic, epidemic, vaccine, GSA, virus, health, disease, infected, booster, death
Topic 2: Social	social, distance, isolation, lockdown, migration, remote, online, curfew, mask, sanitizer
Topic 3: Research	dataset, measures, count, model, analysis, prediction, simulation, optimization, spread, results
Topic 4: Generic	approach, crisis, initiatives, education, precaution, spread, resilience, transport, efficiency, paper

Table 1: Group-wise top ten topic words.

4.3 Sentence Generation

For generating sentences (information) on a similar set of topics, we select the *COVID-19 Open Research Dataset* (CORD-19) (Wang et al., 2020) public corpus as our test set. CORD-19 comprises a comprehensive collection of scholarly articles related to COVID-19. With over 200,000 articles encompassing various disciplines, it serves as a gold standard corpus for NLP and biomedical research. The dataset contains 84 million words and 16 million tokens and facilitates diverse analyses, from topic modeling to information extraction. CORD-19's metadata covers authors, affiliations, and publication dates for each included paper as well.

In our experiment, the topic words from four topic groups extracted from the training set are the input for the primary masking phase of the transformer. These topic words are masked prior to the input by encoding the Context Scores with the words for the awareness of the model. Using the decoder layer, these explicitly masked tokens serve as the base prompts for our Tokenized Generative Transformer model.

The masking technique is important for processing sequences of different character lengths from the topic words while preserving the advantages of parallel processing. The self-awareness mechanism uses attention scores from queries and keys to compute weighted sums of values to capture sequence relationships.

We utilize the GPT-3 language model as the platform for developing the Tokenized Generative Transformer model. After masking the topic words, the model imposes the self-attention mechanism for normalizing the Context Scores for further computation. Accuracy should decrease as the model generates sentences that further deviate from the topic words, indicating that consistency can still be improved. We use self-attention to determine the attention scores for each topic word. We perform masking for self-attention using three trained linear projections: Query (Q), Key (K), and Value (V). The attention scores are generically calculated as follows:

$$\textit{Self-Attention(Q,K,V)} = \textit{softmax}\frac{(QK^T)}{(\sqrt{d_k})} \cdot V \quad (7)$$

where d_k is the dimension of key vectors, or, as in our experiment training set containing the topic words.

The multi-head attention mechanism computes multiple attention scores in parallel, allowing the model to attend to different parts of the input sequence simultaneously. The outputs of the attention heads are concatenated and linearly transformed to produce the final output:

$$\textit{Multi-Head(Q,K,V)} = \sum_{i=1,\ldots,n}^{N} \text{Head} : (h_1, \ldots, h_n) \cdot W^O \quad (8)$$

In Eq. 8, *W* is the weights' sum. $head_n = Attention(Q.W^{(Q_i)}, K.W^{(K_i)}, V.W^{(V_i)})$, $Q.W^{(Q_i)}, K.W^{(K_i)}, V.W^{(V_i)}$, and W^O are trained linear projections from the embedded layer. This sub-layer of linear projection is a FeedForward Network (FFN), which consists of two linear transformations with a ReLU activation function in between:

$$FFN(x) = \max(0, X \cdot W_1 + b_1) \cdot + \ldots + (W_n + b_n) \quad (9)$$

The W_1, W_2, b_1 and b_2 are parameters for the feed-forward layer. Also, since the transformer architecture does not have any inherent notion of the position of words in a sequence, positional encoding is inherently added to the input embeddings to provide information about the relative positions of words.

To evaluate the quality of generated sentences, we first calculate the accuracy of the 100 sentences generated for each topic using Eq. 10. For that purpose, an extended list of topic words is manually created, for each topic word which is consistent with the topic domain. Then we check, within the generated 100 sentences for a topic, what percentage contain topic words from the corresponding extended list. Here T_i represents topic *T1* to *T4*.

$$\textit{Topic-Acc}(T_i) = \frac{(Original_{Sentences}\ T_i)}{(Generated_{Sentences}\ T_i\ \textit{(100)})} \quad (10)$$

The cumulative mean accuracy score is obtained from the *Topicwise_Accuracy* values of all the sentence generation cases.

$$C\bar{M}_{Accuracy} = \frac{\sum_{i=1}^{n} Topic\_Acc(T_i)}{4} \quad (11)$$

The cumulative mean accuracy is obtained for all the generated sentences for topic *T1* to *T4* in terms of simultaneous sentence generation per topic.

5 Results and Analysis

Each topic word acts as a cue for the transformer model to generate 10 sentences with a length of 15 to 25 tokens (words) with a temperature of 0.2. We locally measure the generation accuracy for each generated sentence. The accuracy is calculated using Eq. 10. This accuracy is calculated for all the sentences of each topic. Finally, the cumulative mean of all these topic-wise sentences is computed. In case any generated sentence does not contain any topic word, that sentence is discarded. After the sentence generation on 4 topics, our model achieves a cumulative mean accuracy of **89.54%** in generating sentences, demonstrating the model can generate consistent and fluent sentences from the topic words.

In Table 3, a single generated sentence per topic word is shown from a few topic words already mentioned in Table 1. The generation is also crucial for scrutinizing the sentence styles and semantics produced by the proposed Tokenized Generative Transformer.

5.1 Information Validation Parameters

For comparing the document D_1 consisting of original sentences and D_2 containing generated sentences, we investigate the accuracy and generics of the generated sentences with the actual sentences on the same topics. It has already been shown that in unsupervised and semi-supervised approaches, pre-trained word embeddings are replaceable by contextualized word representations (Peters et al., 2018). We already have masked contextual topic words for higher contextuality representations. These word representations can be better pairwise feature scores for semantic evaluation.

To fully utilize the semantic embedding, we propose a *Perplexity-Similarity Score* to achieve the similarity between the comparing documents to understand the complex similarity or polarity structure of the sentences within. Perplexity scaling is essentially useful for diversified learning, as well as for the evaluation metrics of large corpora. For efficient contrastive selection of sentences with similar perplexities, we define an empirical approach as:

$$Perplexity = P_{O(s)} - P_{G(s)} \quad (12)$$

Index	Topic Word	Generated Sentence Corresponding to Each Topic Word
		Topic T1: Medical Topic; Topicwise Accuracy: 0.92
1.	pandemic	**Pandemic** is a phenomenon when a large number of people are infected with the virus.
2.	epidemic	The **epidemic** is the number of large contaminations that are detected in a given period.
3.	vaccine	Pfizer was the first coronavirus **vaccine**.
		Topic T2: Social Topic; Topicwise Accuracy: 0.89
1.	social	**Social** is a term that describes the social interaction of individuals.
2.	distance	The **distance** between two points on a two-dimensional coordinate is Euclidean.
3.	isolation	WHO reports **isolation** as one of the core reasons for depression.
		Topic T3: Research Topic; Topicwise Accuracy: 0.90
1.	dataset	Many COVID-19 related **datasets**, tools, software are created and shared.
2.	measures	**Measures** is the number of steps that can be taken to achieve the desired outcome.
3.	count	Daily **count** of the COVID-19 cases was at an all-time high in the first quarter of 2020.
		Topic T4: Generic Topic; Topicwise Accuracy: 0.85
1.	approach	The **approach** is a simple, straightforward, and cost-effective way to reduce the cost.
2.	crisis	A setback is not a **crisis**, but scope for analytical examination.
3.	initiatives	All the necessary **initiatives** have been taken to slow down the rate of transmission.

Table 2: Topic-wise information generation. The topic words within generated sentences are marked in distinct **colors** for identification purpose.

P is computed as the perplexity of the comparable sentences to the differences in context scores between the actual information (sentences) $(P_{O(s)})$ and generated information $(P_{G(s)})$.

The comparison criteria for each sentence within D_1 and D_2 is cumulated and determined with each iteration up to *n* by a comparable sentence count θ, concerning the perplexity comparison between the sentences, by deducing the *Hadamard product* (Horn, 1990) of temperature and length of each sentence to be compared one to the same linear dimension.

Once the documents are represented as lists for comparison, we perform the multi-document summarization for encoded tokenized representation of the comparable documents. This is simply done because of optimization and efficiency, to tackle the bottleneck of the system while comparing numerous sentences at each step:

$$E = Comp(\theta^{(D_1 \rightarrow D_2 \ldots)}) \times (s(T \rightarrow t_1, \ldots, t_n)) \quad (13)$$

The encoding scheme takes the comparability factor of each document while fragmenting the generated sentences $(D_2(s))$ and original sentences $(D_1(s))$ into a possible number of elements of the token set *T*. We compute sentence embeddings from comparisons by multiplying encoded tokens. This combines perplexity with pairwise distances between documents to produce a final semantic similarity score, indicating document affinity.

Given two sets of sentences, S_1 and S_2, at first, these sentences are encoded into a numerical format. Let τ be the tokenization function that maps a sentence *s* to a sequence of tokens $\tau_1, \tau_2, \tau_3, \ldots, \tau_n$. Also, let *E* be the encoding function that embeds by mapping each token τ_i to a high-dimensional vector. The encoded input for any sentence *s* can be represented as a matrix *X* of size $n \times t$, where *n* is the number of tokens and *t* is the dimension of the token embeddings. At this step, if any similar sentences are matched from the perplexity perspective, the comparable resultant sentences are printed with their acquired scores. Formally, the Perplexity-Similarity Score can be proposed:

$$\textit{Per-Sim Score} = E \cdot \tau(D_1(s) \approx D_2(s)) \quad (14)$$

The encoded inputs are passed through learnable parameters from the TGT, such as the adaptive self-attention, multi-head attention, and feedforward layer with positional encoding with normalization.

By computing the pairwise variable distance between the embeddings, we can measure the similarity range between the sentences and gain insights into the underlying structure and content of the text. For a better understanding of per-topic comparison metrics within documents D_1 and D_2 from Perplexity-Similarity Score-driven document

Models	ROUGE		METEOR		BLEU		Per-Sim Score	
	Train	Test	Train	Test	Train	Test	Train	Test
T5-11B (Raffel et al., 2020)	0.75	0.72	0.80	0.78	0.85	0.83	0.85	0.85
LLaMA-13B (Touvron et al., 2023)	0.78	0.76	0.82	0.80	0.85	0.82	0.87	0.86
MPT-7B (Team, 2023)	0.76	0.74	0.81	0.78	0.83	0.81	0.84	0.83
GPT Neo-20B (Black et al., 2021)	0.79	0.77	0.83	0.81	0.87	0.85	0.89	0.89
CoLDA+TGT (Ours)	**0.82**	**0.80**	**0.86**	**0.84**	**0.87**	**0.86**	**0.93**	**0.91**

Table 4: Benchmarking of our proposed framework with transformer-based state-of-the-art language models on the open corpus CORD-19.

comparison for each iteration set, we show the classification report in Table 3.

Topic Groups	Acc.	Prec.	Rec.	F1-Score
T1	0.92	0.90	0.90	0.91
T2	0.89	0.91	0.89	0.89
T3	0.90	0.86	0.85	0.90
T4	0.85	0.85	0.85	0.87
Overall	**0.89**	**0.88**	**0.87**	**0.89**

Table 3: Classification report derived from the identification samples.

5.2 Benchmark Evaluation

Dedicated multi-domain NLP tasks have been performed successfully with several transformer architecture-based models, with state-of-the-art results. We evaluate the performances of a few SoTA models on our test data.

For benchmarking, a few language models with comparable parameter sizes[6] are selected for comparison, as the *Unified Text-to-Text Transformer* (T5-11B) (Raffel et al., 2020), *LLaMA* by Meta (13B parameters) (Touvron et al., 2023), *MosaicML Pretrained Transformer* (MPT) (7B parameters) (Team, 2023), and *GPT Neo* (20B parameters) (Black et al., 2021).

In Table 4, the performance of these models is shown on our test data, alongside our proposed framework's overall attained performance. Our framework maintained a stable outcome across the metrics for both the train and test cases, while outperforming the other compared models. The factual similarity between the generated and demonstrated sentences containing the same topic words is also shown in *Per-Sim Score*, where our framework demonstrates a significant enhancement over the other evaluated models. The GPT Neo scores a narrowly followed score in comparison to our framework, indicating the underlying similar architectures (GPT-based) for both the language models.

The superior performance of our framework indicates its ability to generate sophisticated text while preserving factual accuracy. The contextual topics from Co-LDA enable the capture of nuanced contextual relationships. Our tokenizer-based masking technique aids TGT in producing coherent and semantically consistent sentences. As a result, the system pipeline combining these techniques produces an end-to-end method for enhanced performance on factual text generation tasks, demonstrating both fluency and informative correctness.

6 Conclusion

This paper proposes a framework for generating accurate and relevant information from large corpora of text, combining a contextual topic modeling algorithm (Co-LDA) with a tokenizer-based generative transformer (TGT) network. Co-LDA captures the contextual relationships between topics in a corpus, while TGT harnesses the power of state-of-the-art language models to generate informative sentences based on extracted topic words. Experimental findings overcome the limitations of fixed-length sentence comparisons by considering variable-length sentences in training and test cases. Evaluation using several performance indicators and standard validation metrics ensures an informed assessment of the framework's effectiveness and enables meaningful comparisons with existing gold-standard benchmark datasets. Our approach can be adapted and utilized in other domains, facilitating the extraction and generation of impactful insights from high-volume text data.

[6] **B** mentioned with the models stands for parameters sizes in Billions.

References

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow. *If you use this software, please cite it using these metadata*, 58.

D. M. Blei. 2001. Latent Dirichlet Allocation, Advances in Neural Information Processing Systems. *NIPS'01*.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation.

Roger A Horn. 1990. The hadamard product. In *Proc. Symp. Appl. Math*, volume 40, pages 87–169.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. *arXiv preprint arXiv:2304.03245*.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the risk of misinformation pollution with large language models. *arXiv preprint arXiv:2305.13661*.

Letian Peng, Zuchao Li, and Hai Zhao. 2023. Contextualized semantic distance between highly overlapped texts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10913–10931.

Matthew E Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Publisher: OpenAI.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5220–5255.

MosaicML NLP Team. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms. Accessed: 2023-05-05.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, \Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, et al. 2020. Cord-19: The covid-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Xiaobao Wu, Anh Tuan Luu, and Xinshuai Dong. 2022. Mitigating data sparsity for short text topic modeling by topic-semantic contrastive learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2748–2760.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Statistical Measures for Readability Assessment

Mohammed Attia
Google LLC,
US
attia@google.com

Younes Samih
IBM Research,
UAE
younes.samih@ibm.com

Yo Ehara
Tokyo Gakugei University,
Japan
ehara@u-gakugei.ac.jp

Abstract

Neural models and deep learning techniques have predominantly been used in many tasks of natural language processing (NLP), including automatic readability assessment (ARA). They apply deep transfer learning and enjoy high accuracy. However, most of the models still cannot leverage long dependence such as inter-sentential topic-level or document-level information because of their structure and computational cost. Moreover, neural models usually have low interpretability. In this paper, we propose a generalization of passage-level, corpus-level, document-level and topic-level features. In our experiments, we show the effectiveness of "Statistical Lexical Spread (SLS)" features when combined with IDF (inverse document frequency) and TF-IDF (term frequency–inverse document frequency), which adds a topological perspective (inter-document) to readability to complement the typological approaches (intra-document) used in traditional readability formulas. Interestingly, simply adding these features in BERT models outperformed state-of-the-art systems trained on a large number of hand-crafted features derived from heavy linguistic processing. In analysis, we show that SLS is also easy-to-interpret because SLS computes lexical features, which appear explicitly in texts, compared to parameters in neural models.

1 Introduction

A large number of readability formulas (also called shallow readability indicators) have been developed since the 1940's, but most of them use superficial intra-sentential information (e.g., average sentence length and average character length) without using inter-sentential information such as document-level, corpus-level and topic-level statistics.

To address this issue, we introduce Statistical Lexical Spread (SLS), and combine it with features derived from IDF and TF-IDF to train neural and non-neural models on automatic readability assessment (ARA). This set of data-driven features can be extracted from any corpus, preferably where documents are categorized into topics. In this project we utilize Wikipedia where articles, by design, are grouped into categories.

We use these features to augment a BERT-Based classifier on some benchmark data sets to determine if any significant improvement can be gained from these features or they are already learned by BERT embeddings. For this purpose we develop a 'single-shot' model where BERT is fine-tuned on the text alone or text combined with the numerical values of our features, and a 'hybrid' model, where the BERT pipeline is augmented with the predictions from a non-neural classifier. Interestingly, the 'hybrid' mode shows remarkable improvement on the results of the 'single-shot' mode, and overall our models outperform (or compete with) state-of-the-art methods that rely on heavy linguistic processing.

To test the generalizability and crosslinguality of our methods, we evaluate our models on English, Spanish, and Catalan. The advantage of our approach is that no sophisticated NLP processing tools or resources are needed, apart from an optional lemmatizer, which makes it suitable for low-resourced languages. In case a lemmatizer does not exist, SLS+TF-IDF can still be narrowed down to statistics on the surface forms with even better performance on some data sets, while on others yielding only 1.43% absolute below the highest scores.

2 Related Work

With the recent advancement of machine learning (ML), researchers started to apply it to ARA usually modeling it as a classification task. Early studies introducing ML to ARA developed hand-crafted features extracted mostly from the linguistic analysis of texts. For example, Schwarm and Osten-

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 153–161
December 1-3, 2023 ©2023 Association for Computational Linguistics

dorf (2005) introduced four syntactic features (average parse tree height, and average number of noun phrases, verb phrases, and SBARs) to train an SVM classifier. Pitler and Nenkova (2008) enriched that with features indicating lexical cohesion (e.g. number of pronouns in a sentence, and word overlap between sentences), and discourse connectivity (e.g. whether connectives between sentences are implicit or explicit). Over the years, the number of linguistic features kept growing reaching 155 (Vajjala and Lučić, 2018) and 255 (Lee et al., 2021).

Although the focus on generating sophisticated linguistic features might help advance the state of the art for English, it does not generalize well, due to the fact that many languages have limited NLP tools and resources. For example, (Imperial, 2021) used 155 linguistic features for English and only 54 for Filipino due to this limitation. This is why we introduce frequency-based features with minimal NLP tooling requirements (only a lemmatizer) that scales well across languages. We also show that even with the lack of a lemmatizer, the non-morphology based features can still deliver a comparable performance.

There have been a few attempts to depart from linguistic features for ARA, particularly using the help of language models. For example, Collins-Thompson and Callan (2004) developed 12 language models matching the 12 American grade levels. Their language models are unigrams and assume that the probability of a token given the grade level is independent of the surrounding tokens. Cha et al. (2017) used Brown clustering which aims to maximize the mutual information of word bigrams. Language models, however, focus on corpus-level information, and they do not have a mechanism to account for passage-level, document-level or topic-level information, which TF-IDF and readability formulas, for example, prove to be more suited for.

3 Data Collection and Sampling

As is the case with document indexing in Information Retrieval and Text Mining, for the construction of data-driven features, we need to compute weights for each word to quantify the degree of its familiarity. Instead of using a set of web pages, we use Wikipedia articles for our indexing purposes. There are three primary advantages of Wikipedia for our approach to ARA: first, it is available in many languages, second, it covers a broad variety of topics, and third, most articles are associated with categories, which allows us to cluster articles into their related topics. However, the disadvantage is that Wikipedia articles are edited and reviewed to be of a high quality, and therefore lack the noise and variance common in many other natural text types.

Data point	Count
total titles	6,334,131
total categories	1,347,602
total word count	2,363,334,969
titles with categories	4,049,500
singleton categories	310,997
categories $\in$ 80% of titles	1,161
* titles $\in$ 80% of word count	1,812,671
* categories $\in$ 80% of '* titles'	1,131

Table 1: Topography of titles and categories in the English Wikipedia. *: included in the final selection.

Due to the large number of titles in the English Wikipedia, and the fact that many articles are seed articles, i.e. without any substantial content, we sub-sample the data following Pareto's Principle which states that 80% of consequences come from 20% of the causes. For a total number of 6.3m articles we found that 28.62% of them cover 80% of the word count. By contrast, there are 1.3m categories, and we found that 0.09% of them cover 80% of the titles. The reason that category selection seems to go off the bounds for Pareto's Principle is that categories are very liberally used in Wikipedia. For example, 23.08% of the categories are singletons, i.e. representing only one article. Statistics for the English Wikipedia data dump of September 2^{nd}, 2021 are shown in Table 1. The same sub-sampling strategy is used for the other two languages tested in this project, i.e. Spanish and Catalan.

4 Feature Design and Selection

Understanding a document is dependent on the reader's level of familiarity with the underlying knowledge base (or the topic), which accounts for the connections in the mental map (Liu and Yuizono, 2020) of the reader and controls the flow of information for updating these connections. This underlying knowledge base indicates the presence or absence of the shared world knowledge between the writer and the reader. Approximating this underlying knowledge map can be obtained by analyzing the connection between words within a document and across a reasonably large collection of documents.

Topic modeling for ARA has been discussed in a number of papers. For example, Qumsiyeh and Ng (2011) developed a system called ReadAid that used Latent Dirichlet Allocation (LDA), an unsupervised learning algorithm, to determine the ranked probability of topics covered in a document based on the distribution of words in the document, or more precisely the probability of a word given a topic $P(w|t)$, and the probability of a topic given a document $P(t|d)$. Their model was trained on 53 subject areas extracted from the English Curriculum and College Board[1] and a set of 100 documents randomly selected for each subject area from DMOZ[2]. Lee et al. (2021) expanded this approach by training LDA models on four variations of 50, 100, 150, and 200 topics, and analyzed the output for semantic richness, clarity, and noise, with the purpose of understanding how the topics are distributed, not just what they are.

In this work we use SLS+TF-IDF to model topics using top-frequency categories already curated in the Wikipedia corpus, with the intuition that specialized words will occur in fewer categories than common ones. The categories selected in our analysis are based on a ratio of titles and word counts as explained in Section 3, and the number is 1,131 for English, 1,536 for Spanish, and 1,516 for Catalan.

The features used in this project are divided into four groups: Statistical Lexical Spread (SLS), TF-IDF features, document-based counts, and traditional readability formulas (RF). Details are explained in the following sub-sections.

4.1 Statistical Lexical Spread (SLS)

The main intuition for SLS is that easy words occur more often and in more contexts, spanning more articles and more topics, than difficult words. Even if a word is long and multi-syllabic, such as 'television', if it occurs more often, it will be considered more readable than less frequent words, even if they are short and monosyllabic, such as 'deuce'. Another intuition is that words which show a high morphological variability, such as 'play, plays, played, playing', are generally easier to read than rigid and uninflected words, such as 'timid'. The advantage in SLS is that frequency statistics are gathered at the corpus level, document level and topic level.

For the three features of 'unknown_word', 'uninflected_word' and 'below_mean_count' we just take the ratio (count of positive tokens divided by the total number of tokens in a document). For the other seven features we take the log of the average according to Equation 1, where t is a term which can be a lemma or a form, $f(t)$ is the function that retrieves the frequency, spread or variability value, and $l(d)$ is the length of the document. Features with the suffix '$\_freq$' are for corpus-level statistics, '$\_article\_spread$' for document-level statistics, and '$\_category\_spread$' for topic-level statistics. For English lemmatization, we use NLTK (Bird et al., 2009), and for Spanish and Catalan, we use spaCy (Honnibal and Montani, 2017).

$$\log\left(\frac{\sum_{t=1}^{l(d)} f(t)}{l(d)}\right) \tag{1}$$

Non-Morphology Features:
1. form_freq: form frequency, or how many times a form occurred in the entire corpus.
2. form_article_spread: in how many articles a form appeared, regardless of total frequency.
3. form_category_spread: in how many categories a form appeared.
4. unknown_word: words that do not occur in the corpus or have a frequency below a certain threshold, which is set in our experiment to 10.
5. below_mean_count: words that have a frequency below the mean frequency of the word list consumed (excluding unknown words above). This happens to be 1441.79 in the English Wikipedia sample.
Morphology-based Features:
6. lemma_freq: lemma frequency, or how many times a lemma occurred in the entire corpus.
7. lemma_article_spread: in how many articles a lemma appeared, regardless of the total frequency.
8. lemma_category_spread: in how many categories a lemma appeared.
9. morph_variability: for each lemma, how many different forms are represented by the given lemma. This is an indication of morphological richness.
10. uninflected_word: words that do not have any morphological inflection in the corpus.

The use of the log in the calculations is meant as a normalization step to dampen the effect of exploding numbers when the numerator is much greater

[1] www.collegeboard.org
[2] www.dmoz-odp.org

than the denominator and the variance between different outputs cannot fit in a scale.

4.2 TF-IDF Features

TF-IDF has been used in the readability literature in two different ways.

1. Using the TF-IDF for all tokens in a given document as a vector (Chen et al., 2011).
2. Using the mean of TF-IDF of all tokens a document (De Clercq et al., 2014).

TF-IDF is powerful in collecting statistics on term distribution and weight across a collection of documents. In our research, we use the mean of TF-IDF and the mean of the IDF for forms and lemmas in a document. This gives us 4 features. We further apply it to articles, and categories as documents (for topic modeling). This expands the number of features to 8, and this allows us to utilize the power of the TF-IDF orthogonally at the document level and the topic level. Equations 2, 3, and 4, show how the calculations are conducted, where t is the term which can materialize as a form or a lemma, d is a document, D is a collection of documents (or categories), l is the length function, c is the counting function, e.g. $c(t, d)$ is the count of term repetitions in a given document, and uc is a unique counting function, i.e. if a term occurs in a document one or more times, it will be reduced to one, otherwise zero.

$$TF_{(t,d)} = \frac{c(t,d)}{l(d)} \tag{2}$$

$$IDF_{(t,D)} = \log\left(\frac{l(D)}{uc(t,D)}\right) \tag{3}$$

$$TF\text{-}IDF_{(t,d,D)} = TF_{(t,d)} \times IDF_{(t,D)} \tag{4}$$

Here we list the features derived from TF-IDF also divided into whether they are dependent/non-dependent on morphological analysis (lemmatization).

Non-Morphology Features:

1. form_article_idf: average IDF where t is a word form and D is a collection of articles.

2. form_category_idf: average IDF where t is a word form and D is a collection of categories.

3. form_article_tf-idf: average TF-IDF where t is a word form and D is a collection of articles.

4. form_category_tf-idf: average TF-IDF where t is a word form and D is a collection of categories.

Morphology-based Features:

5. lemma_article_idf: average IDF where t is a word lemma and D is a collection of articles.

6. lemma_category_idf: average IDF where t is a word lemma and D is a collection of categories.

7. lemma_article_tf-idf: average TF-IDF where t is a word lemma and D is a collection of articles.

8. lemma_category_tf-idf: average TF-IDF where t is a word lemma and D is a collection of categories.

4.3 Document-Based Features

We need to account for passage-level information, such as a word repetition, word count and the type of lexicon used. These features are computed locally by counting words in a given document, or matching them against predefined lists.

1. word_count: word count in the current document, taken as a ratio against the maximum word count found in a document set.

2. word_rep: in a given document, how many times a word is repeated. This is then averaged against total words in a document

3. basic_vocab: how many words are found in a list of basic vocabulary. The source of the word list is Simple Wikipedia list of 1000 basic words. This is taken as a ratio against total words in a document.

4.4 Readability Formulas (RF):

Readability Formulas (RF) are known for their efficiency at capturing passage-level information. There are a few python implementations of these formulas. In this project, we chose the implementation in TextStat.[3] Here is a list of the formulas used:

1. Flesch Reading Ease, (Kincaid et al., 1975).
2. Flesch-Kincaid Grade, (Kincaid et al., 1975).
3. SMOG Index, (Mc Laughlin, 1969).
4. Coleman-Liau Index, (Coleman and Liau, 1975).
5. Automated Readability Index, (Smith and Senter, 1967).
6. Dale-Chall Readability Score, (Dale and Chall, 1948)
7. Linsear Write Formula[4].
8. Gunning-Fog Index, (Gunning et al., 1952).
9. Text Standard, based on consensus among a number of tests.

[3] https://github.com/textstat/textstat

[4] https://en.wikipedia.org/wiki/Linsear_Write

10. Fernandez-Huerta, (Fernández-Huerta, 1959).
11. Szigriszt-Pazos, (Szigriszt Pazos, 1992).
12. Gutierrez Polini, (Gutiérrez de Polini, 1972).
13. Crawford, (Crawford, 1985).
14. Gulpease Index[5].
15. Osman, (El-Haj and Rayson, 2016).
16. Difficult Words, (Dale and Chall, 1948).

4.5 Feature Subsets

For some ML algorithms, the high dimensionality of features can be problematic. Therefore, we use XGBoost to determine the important features based on training on the English Viki-Wiki data set. We select the overlap between gain and coverage, and here are the subsets selected.
Selected_8 = (4.1): 4, (4.1): 2, (4.4): 8, (4.4): 6, (4.2): 4, (4.2): 1, (4.4): 7, (4.3): 1
Selected_6 = first 6 in selected_8.
SLS_8 = (4.1): 6, (4.1): 2, (4.1): 1, (4.1): 7, (4.1): 9, (4.1): 4, (4.1): 3, (4.1): 8
SLS_6 = first 6 in sls_8.
RF_8 = (4.4): 8, (4.4): 6, (4.4): 7, (4.4): 11, (4.4): 2, (4.4): 3, (4.4): 13, (4.4): 15
RF_6 = first 6 in rf_8.

5 Correlation with Readability Formulas

We conducted a comparison between our statistical measures and the traditional readability formulas to see to what degree they are aligned on their predictions. The data set used in this experiment is the Simple Wikipedia, with a total number of instances of 142,759. We used a split of 66% for training and 34% for testing. We applied the decision tree Random Forest algorithm, and the results are shown in Figure 1.

Generally, there seems to be a strong correlation between our SLS and dale_chall_readability_score, while document-based and TF-IDF have the highest correlation with ‘difficult_words’. We notice that some non-English specific indicators, such as ‘osman’, ‘gulpease_index’ and ‘gutierrez_polini’ have higher correlation with our criteria than some English-specific ones, such as ‘gunning_fog’ and ‘smog_index’. This is why we decided to use all 15 formulas in subsequent experiments.

It’s also interesting to consider the correlation coefficient among the different traditional readability formulas. Many pairs of formulas have high correlation, whether negative or positive, which means that they are looking at the same or similar pieces of information, while many other pairs have a correlation between -0.5 and 0.5 which indicates low correlation, meaning they are looking at different pieces of information, or interpreting the same pieces of information differently.

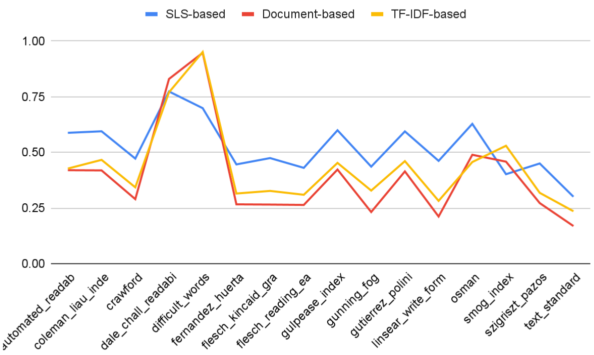


Figure 1: Correlation between new and traditional readability formulas

6 Testing on Benchmark Test Sets

We use our features, along with the readability scores from the traditional readability formulas and build ML models and apply them to two benchmark data sets: a monolingual one, OneStopEnglish, and a multilingual one, VikiWiki.

6.1 OneStopEnglish (OSE)

OneStopEnglish (Vajjala and Lučić, 2018) is a collection of articles obtained from the Guardian newspaper and adapted by teachers for three levels of learners (elementary, intermediate, and advanced). The data set contains 564 instances (189 elementary, 189 intermediate, and 186 advanced.

6.1.1 Classification with Non-neural Classifiers

Table 2 shows the results of the experiments with 10-fold cross validation using a number of non-neural ML classifiers. Our best models give an accuracy of 80.15% using the ‘selected_8_no_morph’ features in an SVM classifier. This outperforms the results in (Vajjala and Lučić, 2018) which was 78.13% using 155 linguistic features, and the results in (Lee et al., 2021) which was 77.8% using 255 handcrafted features.

In our initial experiments, we noticed that the results for train-test splits can vary dramatically by the split size, while n-fold cross validation gives

[5] https://it.wikipedia.org/wiki/Indice_Gulpease

Feature sets	LR	SVM	XGB	RF
SLS	43.24	38.97	35.42	31.54
Doc-based	65.58	69.51	67.72	66.67
TF-IDF	66.11	68.45	64.71	63.84
RF	69.85	74.82	75.19	74.12
selected_8	72.16	79.26	77.31	78.38
selected_8 _no_morph	72.15	**80.15**	77.14	77.66
selected_6	69.14	78.02	78.37	77.48
SLS_8	42.52	39.69	34.88	33.85
SLS_6	39.70	37.38	35.61	32.92
RF_8	59.58	60.65	57.27	58.88
RF_6	59.22	60.47	55.85	57.79
All features	73.93	77.48	75.52	77.66

Table 2: ML Classification results on OSE. LR = Logistic Regression, XGB = XGBoost, RF = RandomForest

more reliable results, particularly for smaller data sets.

6.1.2 Classification with BERT Fine-Tuning

Since its inception, BERT (Devlin et al., 2018) has shown strong performance on many NLP benchmark data sets. There were a number of attempts to apply it to ARA, including that of Martinc et al. (2021), who reported an accuracy of 67.38% training on text alone without additional features. Their best result on the OSE data set was 78.72% using HAN (Hierarchical attention networks).

Combining BERT embeddings with additional features has been explored in a number of papers and most of them used the fused features in a non-neural classifier (Deutsch et al., 2020). For example, (Imperial, 2021) extracted BERT embeddings, concatenated them with 155 linguistic features, making a total of 923 dimensions, and fed that into a number of ML classifiers. Imperial (2021)'s best result was an F1 score of 73.2% using logistic regression.

In a more recent paper, (Lee et al., 2021) reported 80.1% mean accuracy on five-folds on OSE using BERT without handcrafted features. They further managed to increase the accuracy to 98.2% using a hybrid model, where they took the predictions of BERT fine-tuning, along with 255 handcrafted features, and fed them to a non-neural classifier.

In this paper, we experiment with BERT in two modes: 'Single-Shot Mode' and 'Hybrid Mode', as explained below. For English we use the model 'bert-base-uncased' for English and 'bert-base-multilingual-uncased' for the other languages.

Parameters	values
Epochs	16
Learning rate	1e-5
max token length	200
batch size	32
optimizer	AdamW
number of folds	10

Table 3: BERT classification setup.

Single-Shot Mode: Here we combine our features with the text embeddings following Chris McCormick article on "Combining Categorical and Numerical Features with Text in BERT".[6] We use the model 'transformers.BertForSequenceClassification' with the parameters listed in Table 3.

We tried two ways of appending the numerical values of features to the text. The first was to include the numerical values separators, and the second was to concatenate the feature name along with the numerical value. We found that the second method worked best, and this is what is reported in this paper.

1. `f'{value} [SEP]. '`
2. `f'{feature}: {value} [SEP]. '`

Hybrid Mode: Similar to Lee et al. (2021), we also build a hybrid model, but instead of using BERT predictions in a non-neural model, we use the predictions of a non-neural model and feed them to the BERT fine-tuning along with the feature sets and the text embeddings. We first train SVM on 'selected_8_no_morph', take the predictions for each fold (so that there is no chance for over-fitting), and combine them together as an additional feature in BERT fine-tuning.

Results for both the single-shot and the hybrid model are shown in Table 4. All experiments are conducted with 10-fold cross-validation. Our baseline is BERT fine-tuned on text embeddings only without any features, which is 86.16% for the single-shot model. This is significantly greater than the 80.1% reported by Lee et al. (2021). The best result for the single-shot mode was 96.64% when BERT embedding is concatenated with the

[6] https://mccormickml.com/2021/06/29/combining-categorical-numerical-features-with-bert/

Features combined	Single-Shot	Hybrid
No features used	86.16	93.42
All features used	39.54	77.84
SLS	29.09	77.84
Doc-based	75.55	91.84
TF-IDF	**96.64**	**98.23**
TF-IDF_no_morph	88.51	96.80
RF	70.94	82.63
Selected_8	77.83	97.52
Selected_6	86.31	95.56
SLS_8	87.42	98.05
SLS_6	87.01	95.38
RF_8	81.72	91.12
RF_6	79.61	84.56

Table 4: BERT fine-tuning results on OSE.

eight features of TF-IDF. For the hybrid mode, our best result is 98.23%. We notice that the hybrid mode gives a significant boost to the performance on most of the features used.

It must be noted that (Lee et al., 2021) managed to obtain 96.5%, and 96.8% accuracy in single-shot mode using RoBERTa and BART respectively, and 99.0%, and 97.1% in a hybrid mode.

6.2 Viki-Wiki

VikiWiki is a multilingual readability data set of Vikidia articles and their Wikipedia counterparts (Madrazo Azpiazu and Pera, 2020). Vikidia[7], like Wikipedia, is an encyclopedic website providing information on various topics in English and a number of other languages, but the main goal is to make the content simple and easy to read. The number of instances in the dataset is 864 for English, 831 for Spanish, and 855 for Catalan. The best results reported by Madrazo Azpiazu and Pera (2020) (in terms of accuracy for 10-fold cross-validation) was 96% for English, 87% for Spanish and 96% for Catalan. In their work, Madrazo Azpiazu and Pera (2020) used different sets of features including shallow, morphological, syntactic, and semantic features.

6.2.1 Classification with non-neural Classifiers

Table 5 shows the results of our system trained on a combination of features. Our results are comparable to those of (Madrazo Azpiazu and Pera, 2020) for English, Spanish and Catalan. We found that XGBoost gives the best performance compared to other ML algorithms (and this is why we report XGBoost results only here). Again all experiments are conducted with 10-fold cross-validation.

Features sets	en	es	ca
SLS	90.62	82.68	94.38
Doc-based	93.98	84.72	95.56
TF-IDF	92.48	82.32	94.15
RF	95.60	84.61	94.62
Selected_8	95.37	86.65	95.21
Selected_6	**95.95**	86.05	95.09
Selected_6 _no_morph	95.26	84.49	94.63
SLS_8	89.12	81.72	93.57
SLS_6	89.93	81.95	92.05
RF_8	90.74	83.88	93.80
RF_6	90.16	84.37	94.04
All_features	95.37	86.77	**95.79**
All_features _no_morph	95.49	**87.49**	**95.79**

Table 5: XGBoost Classification results on Viki-Wiki.

6.2.2 Classification with BERT Fine-tuning

Following the same approach above with OSE in concatenating numerical values to text in BERT embedding in a **single-shot mode** (Section 6.1.2), we conducted 10-fold cross-validation experiments for English, Spanish and Catalan. The results are shown in Table 6. In most cases the performance converges to 100%. We are not entirely sure about the reason, but it can be due to the fact that BERT is already trained on Wikipedia data, or the task is too easy as the language in the two data sets is clearly distinct. The best published results in the literature is 96% for English and Catalan, which is already high.

Conclusion

For ARA, hand-crafted features derived from heavy linguistic processing do not transfer well across languages, as it becomes harder to find reliable processing tools for low-resourced languages. Our system, by contrast, achieves better or comparable results using only 38 features that capture passage-level, corpus-level, document-level, and topic-level information, and can be computed statistically from any corpus in any language with a light-weight morphological processing tool. We show that even with the absence of a lemmatizer, non-morphological

[7] https://en.vikidia.org

Features sets	en	es	ca
SLS	83.10	80.10	92.84
Doc-based	99.13	100.00	99.88
TF-IDF	100.00	99.88	99.87
RF	99.89	99.71	99.87
Selected_8	100.00	100.00	100.00
Selected_6	100.00	100.00	99.88
SLS_8	100.00	100.00	99.87
SLS_6	100.00	100.00	100.00
RF_8	100.00	100.00	100.00
RF_6	100.00	100.00	100.00
No features	99.89	100.00	99.86
All features	94.42	79.61	90.69

Table 6: BERT Classification results on Viki-Wiki.

features can still yield comparable results. We also show how topic modeling for ARA can be achieved through treating categories as documents in computing features such as IDF and TF-IDF.

Limitations

SLS+TF-IDF provides information on word difficulty and topical specificity drawn from actual language use. One presumed shortcoming of the proposed approach is that it focuses on the lexical statistical behavior and ignores semantic, syntactic and discourse features. Due to the utilization of a lemmatizer, the system can distinguish between 'flag' as a noun and a verb, but it will not be able to distinguish between 'lead' as a metal and or a leash.

Another limitation of the results is that the use of BERT and neural net, by nature, gives different results each run. Although we use 10-fold cross-validation and a relatively higher number of epochs to narrow down the effect of this variability, it is still possible to get slightly different results for each run.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Miriam Cha, Youngjune Gwon, and HT Kung. 2017. Language modeling by clustering with word embeddings for text readability assessment. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2003–2006.

Yaw-Huei Chen, Yi-Han Tsai, and Yu-Ta Chen. 2011. Chinese readability assessment using tf-idf and svm. In *2011 International Conference on Machine Learning and Cybernetics*, volume 2, pages 705–710. IEEE.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Kevyn Collins-Thompson and James P Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of the human language technology conference of the North American chapter of the association for computational linguistics: HLT-NAACL 2004*, pages 193–200.

A Crawford. 1985. Fórmula y gráfico para determinar la comprensibilidad de textos del nivel primario en castellano. *Lectura Y Vida*, 4:18–24.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Orphée De Clercq, Véronique Hoste, Bart Desmet, Philip Van Oosten, Martine De Cock, and Lieve Macken. 2014. Using the crowd for readability prediction. *Natural Language Engineering*, 20(3):293–325.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. *arXiv preprint arXiv:2006.00377.*

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Mahmoud El-Haj and Paul Edward Rayson. 2016. Osman: A novel arabic readability metric.

J Fernández-Huerta. 1959. Medidas sencillas de lecturabilidad [simple readability measures]. *Consigna*, 214:29–32.

Robert Gunning et al. 1952. Technique of clear writing.

L.E. Gutiérrez de Polini. 1972. Investigación sobre lectura en venezuela. *las Primeras Jornadas de Educación Primaria.*

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Joseph Marvin Imperial. 2021. Bert embeddings for automatic readability assessment. *arXiv preprint arXiv:2106.07935.*

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Bruce W Lee, Yoo Sung Jang, and Jason Hyung-Jong Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. *arXiv preprint arXiv:2109.12258*.

Ting Liu and Takaya Yuizono. 2020. Mind mapping training's effects on reading ability: Detection based on eye tracking sensors. In *Sensors (Basel, Switzerland) vol. 20,16 4422. 7 Aug. 2020, doi:10.3390/s20164422*.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2020. Is cross-lingual readability assessment possible? *Journal of the Association for Information Science and Technology*, 71(6):644–656.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 186–195.

Rani Qumsiyeh and Yiu-Kai Ng. 2011. Readaid: a robust and fully-automated readability assessment tool. In *2011 IEEE 23rd International Conference on Tools with Artificial Intelligence*, pages 539–546. IEEE.

Sarah E Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL'05)*, pages 523–530.

Edgar A Smith and RJ Senter. 1967. *Automated readability index*, volume 66. Aerospace Medical Research Laboratories.

Francisco Szigriszt Pazos. 1992. Sistemas predictivos de legilibilidad del mensaje escrito: fórmula de perspicuidad.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

A Question of Confidence: Using OCR Technology for Script analysis

Antonia Karaisl, Waseda University

Abstract

The following article proposes a method employing the Tesseract OCR engine to aid palaeographic analysis and scribal identification. Repurposing the so-called confidence score provided by the OCR engine, different methods of visualization are used to surface differences between font families, script types and manuscript hands.

1 Introduction

This paper introduces a simple method for conducting technology-assisted analysis of script and handwriting styles in printed books and manuscripts. The approach described uses a side product of a tried and tested technology: Optical Character Recognition (OCR). OCR software is traditionally employed for the automatic transcription of text from a digital image to machine-readable output. The method used here largely ignores the transcription but focuses on the so-called confidence scores. Confidence scores are usually employed in the process of OCR recognition to assess the probability that the output is correct. Normally, low-scoring results are undesirable as they signal a lower probability of accuracy. In this case, however, low scores will be used to identify pages, words or characters that could be of interest for a palaeographic analysis.

Digital methods, including Artificial Intelligence (AI), have been plied before to the field of palaeography, for example for the purpose of automatic transcription, the classification of writing styles or scribal identification (see e.g. Camps 2014; Castro Correa 2014; Christlein 2018; Cilia et al., 2019). A contest organized in 2017 by the Fifteenth International Conference on Frontiers in Handwriting Recognition in 2017, for example, solicited AI-based solutions for the classification of medieval script types, providing a set of labelled training material ("ICDAR2017 Competition on the Classification of Medieval Handwritings in Latin Script" 2017). A report submitted by Kestemont et al. discusses the efficacy of several submissions and remarks that the premise of the task itself builds on a simplified reality. Medieval script types do not always have firm boundaries, that is, definitive sets of features that reliably set one type apart from another; some hybrid forms are not easily described with one single label. The categorization of medieval script types often moves in grey zones and a technology trained on human-labelled script types is therefore not automatically free from human bias. Conversely, unsupervised learning, which does not rely on labelled data, does not necessarily sort material in ways that are meaningful to scholars and can also be hard to interpret (Kestemont, Christlein, and Stutzmann 2017, 104–7). Thus, whilst AI shows potential for palaeographic analysis, some caution is warranted: worst case scenario, human subjectivity is replaced by AI's accountability gap.

The approach introduced here is not meaning to replace human expertise with an automated solution. Rather, it attempts to re-purpose a pre-existing technology as a heuristic tool that can accelerate the palaeographer's, book historian's, bibliophile's quest for areas of interest in a book or on a page, with the help of OCR confidence scores. The paper mainly showcases different modes of visualizing confidence metrics to aid the discovery of palaeographic phenomena. The following argument will briefly introduce the metrics of word and character confidence and how they can be employed in script or scribal analysis. Proposed approach employs confidence scores to identify divergences in script style or abnormalities in letter shapes on book- or page level. The elements identified are then inspected with the help of statistics to establish whether they are of significance. Throughout, the open-source OCR

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 162–171
December 1-3, 2023 ©2023 Association for Computational Linguistics

engine Tesseract (LSTM version) is used with different OCR models.[1]

2 Visualizing confidence scores on printed and handwritten material

2.1 Confidence in Theory

In OCR (Optical Character Recognition, typically used for printed text) and HTR (Handwritten Text Recognition) technology, confidence scores are usually procured to aid decision-making calls on transcriptions during the recognition process. By definition, the confidence score marks the probability with which the OCR engine deems the transcription of a character or a word to be correct. Thus, lower confidence scores signal lower probability of accuracy. Confidence is not an absolute measure, and it is possible that low confidence scores can accompany accurate output, or high-confidence scores inaccurate transcriptions. For example, an OCR model trained on predominantly English texts will recognize the same font in a French-language document with fair accuracy but might produce low confidence scores on account of its English-language training. Conversely, a high confidence score does not always guarantee an accurate result – merely the engine's assessment that given the model's parameters the output is accurate.

Tesseract can produce confidence scores at word level and character level. The calculation of the respective confidence scores is complex but understanding the context of their generation can help gauging the underlying parameters. Tesseract's documentation for the current, neural-network-based version (4.0.0 and higher) does not contain any information on the calculation of the confidence score. The documentation for Tesseract's Legacy engine (version 3.0.0 and lower), however, specifies the circumstances and formula for calculating character and word confidence in the context of character classification. When processing new input with the Legacy Engine, Tesseract performs the segmentation of a text image into lines and words down to individual characters. Each segmented character is then classified by mapping it to the closest-matching prototype. The shape of the character is described by a visual feature vector combining a number of 3-dimensional features mapping the character's outline; the distance between the recognized character's visual feature vector and that of the closest matching prototype is then used to calculate the character's confidence (Perveen, n.d.; Smith 2007). On a word-level, the confidence of the lowest-scoring character doubles up as the confidence score for the entire word. (Tesseract Documentation FAQ). If the word formed from the recognized characters turns out highly improbable or linguistically implausible, the Legacy engine tries to re-segment the characters in different ways to see whether a more satisfactory solution can be found (Smith 2007). Output of the Legacy engine expresses character confidence as a percentage; the percentages for all character suggestions for one symbol stand independently and do not necessarily add up to 100%.[2]

In 2016, Tesseract's system was upgraded to include recurrent neural networks with LSTM (Long-Short-Term-Memory). The advantage of such a network is its capacity for context-aware processing, particularly with LSTM, resulting in lower error rates (Ul-Hasan et al 2013). Tesseract's LSTM implementation was adapted from OCRopus, an OCR system based on convolutional neural networks (Smith 2016). The novelty of OCRopus' initial design vis a vis Tesseract's concurrent version was documented at its inception in 2008. In contrast to Tesseract's Legacy engine, OCRopus' recognition process does not segment a text image into separate, single characters, but takes words as base units: proceeding sequentially across an identified word in so-called timesteps, the string is oversegmented – meaning, the word is not chopped into a discrete set of characters, but each timestep presents a separate segmentation attempt for part of the word. Each of these segments presents a character hypothesis; each character hypothesis is assigned a probability that it presents a valid character, and assuming that it is, the "posterior probability" (or confidence) for each character class, based on image features. Recognition results and the relationships between each potential character are expressed in a graph structure, from which the best sequence

[1] Tesseract models and test books used in this study are listed in the appendix.

[2] As noted in the commentary to Tesseract's code, see https://github.com/tesseract-ocr/tesseract/blob/7c178276d78fc4d2e55d531563275fd9631a72fb/src/ccmain/ltrresultiterator.cpp#L458

representing the whole word is then identified, using statistical language modelling (Breuel 2008).

OCRopus' LSTM system is integrated in Tesseract 4.0.0 and later. The last layer of Tesseract's network moreover contains a so-called softmax classifier which normalizes the confidence score for each character before the final output. (Smith 2016). Where Tesseract's Legacy engine yields the confidence as the original percentage in the final output, the LSTM-based Tesseract engine normalizes the confidence score for the chosen character so that the confidences for all classification attempts for one character approximately add up to 1.[3] The final confidence score is amplified through this normalization process, with the result that for the LSTM engine, the character confidences usually move within a band between 1 and 0.90. The softmax normalization only applies to the character confidence, not the word confidence, which fluctuates between 0 and 100%.

Tesseract's OCR models utilize neural networks trained on a pool of so-called ground truth, that is a quantity of labelled material. In the case of OCR, this material consists of image files of text lines, matched with transcriptions saved in simple text files. Throughout the training process, the neural network of the OCR engine iterates over this ground truth, producing transcriptions and evaluating their accuracy against the ground truth labels. Each iteration produces a model which is assessed on its word and character error rate with a separate pool of ground truth. At the end of the process, the model with the highest accuracy is chosen and can then be used with the OCR engine for the automatic transcription of related image material.

The ground truth pool used for training is the key parameter determining the model's capacity to interpret real material – and strategizing on its size and composition is crucial for an effective OCR strategy. In most cases, the aim is to train a model that is specific enough to perform well within its context but capable of generalizing beyond the ground truth pool. Within certain boundaries – say, a language, an alphabet, or a font group – diversity within a ground truth pool can improve this the model's ability to generalize. Too little specialization means the OCR model does not work well within its context. Yet if the ground truth pool is too small or not sufficiently diversified, so-called overfitting occurs: the neural networks' recognition capacity is over-adjusted to its training material. (Kestemont et al. 2017, 97). Since the approach introduced here is less interested in transcription than analysis, some of the models used are deliberately overfitted in order to understand how they can signal affinity or difference between a very small and specific training pool and the test material.

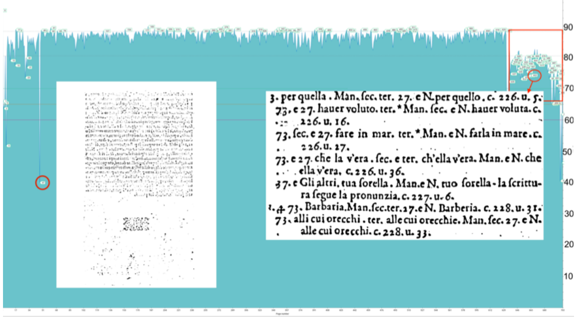
As said, the recognition process employs confidence scores to assess the probability that an output is correct. The confidence scores not only highlight potential good or bad transcriptions; they communicate the efficacy of the chosen OCR model with regard to the input material – and by extension, the affinity of the material it was trained on with the material it is used on. As a consequence, characters underrepresented in a LSTM training set tend to be misclassified in transcription (Ul-Hasan et al 2013). We would expect, therefore, that atypical glyphs or letter shapes not included in the training material are likely to be badly transcribed and consequently flagged up by low confidence scores when running the OCR model over a text image. This is the assumption that the following experiment is seeking to corroborate and utilize.

The caveat to this approach is that the confidence score is an uncertain metric, which is not only affected by the character's shape, but also by factors such as image quality, skewing, or discolorations on the page. With an LSTM-based system, moreover, the impact of context on the confidence score remains difficult to gauge. Overall, therefore, confidence is too complex a metric to serve as an absolute indicator; what the following experiment means to show empirically, however, is that using tools to visualize confidence metrics can still help to identify areas of palaeographic interest on book or page level.

2.2 Confidence in Practice

When analysing the OCR output across a whole book, the word confidence score can help identifying problem areas across the full range using the the average word confidence score for each page in a whole book. Sections where the

[3] See the link to the code commentary in Footnote 2.



3. per quella . Man. ſec. ter. 27. e N. per quello . c. 226. u. 5.
73. e 27. hauer voluto. ter. * Man. ſec. e N. hauer voluta. c. 226. u. 16.
73. ſec. e 27. fare in mar. ter. * Man. e N. farla in mare. c. 226. u. 27.
73. e 27. che la v'era . ſec. e ter. ch'ella v'era. Man. e N. che ella v'era. c. 226. u. 36.
37. e Gli altri. tua ſorella . Man. e N. tuo ſorella - la ſcrittura ſegue la pronunzia. c. 227. u. 6.
4. 73. Barbaria. Man. ſec. ter. 27. e N. Barberia. c. 228. u. 31.
73. alli cui orecchi . ter. alle cui orecchie. Man. ſec. 27. e N. alle cui orecchi. c. 228. u. 33.

Figure 1: Word confidence graph, book-level

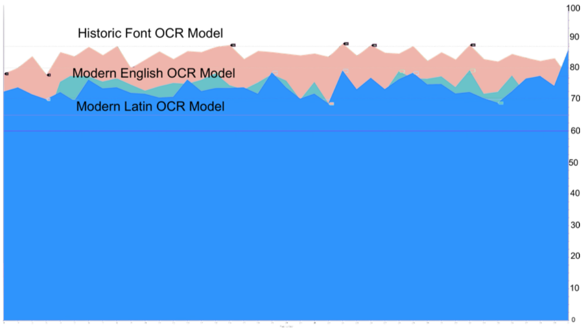


Figure 2: Word confidence graph for different Tesseract models, run over historic printed book

scores move in a narrow, consistent band tend to be transcribed with fair accuracy. Strong fluctuations or exceptionally low confidence scores typically denote pages that differ in some way from the rest. In the word confidence graph for Giovanni Boccaccio's Decameron (see Figure 1), for example, the egregiously low score in the beginning section (circled in red) corresponds with a blank page with ink bleed-through misinterpreted by the OCR model as writing. The finishing section of the same graph, too, shows up with consistently lower scores. Looking at a sample page from this section, it turns out that it contains the book's appendix, which features uncommon symbols, more numbers than usual and truncated words scoring low in confidence, regardless of whether they are transcribed correctly or not (see Figure 1). These particular examples might or might not be of concern. Rather, the point is that outlier pages in OCR output can be identified from the top level with the help of a confidence graph. Beyond the identification of outlier pages, however, there is no further indication what the issue could be in each case – the page image itself needs to be considered to understand what the cause of the low confidence score might be.

Most of Tesseract's standard OCR models are ostensibly trained for specific languages, on modern alphabets (see available list on Tesseract Github). When using these standard models, we would expect low confidence scores (independently from actual accuracy) where an OCR model is plied outside its "comfort zone", so to speak, e.g. to a text containing unknown or underrepresented characters, written in a different language or in a radically unusual font. Granted there is rarely much information about the training material used to train the model (and granted it tends to be too copious for a close review), it is not immediately obvious where such a "comfort zone" starts or ends. Running the model over different kinds of materials and comparing the confidence score, however, can help to get a bearing of the model's capacities. Conversely, running several models over the same material provides some context within which to judge each model's "comfort" and "discomfort zone".

Tesseract's standard models are very effective for modern printed text, and particularly the English language model shows a high performance for many different types of fonts, including typewritten texts. Viewed from the point of methodology, therefore, we should assume that a digitized text's language would be the main parameter to affect confidence scores when it comes to OCR processing. Manuscripts, modern or old, tend to fare badly with these standard models. This is not particularly surprising, granted handwriting tends to be much more irregular than print. Yet experiments with historic printed text, too, show that the OCR models trained for modern languages struggle with such material – as opposed to OCR models trained on a mix of languages, but on historic printed text.

In the following experiment, a set of pages from a Latin publication printed in 1475 were processed with three Tesseract models: the first one trained on English language material printed in modern fonts; the second on Latin language material printed in modern fonts; and the last one trained on material printed between 1500-1800, in Latin, English and French. The accuracy of the transcription was not considered in this experiment; only the confidence scores were assessed in order to understand each model's "comfort" or "discomfort" with historic material printed in Latin. Comparing the graph mapping the average page confidence (see Figure 2), we see that the models for modern English and modern Latin roughly play out in the same

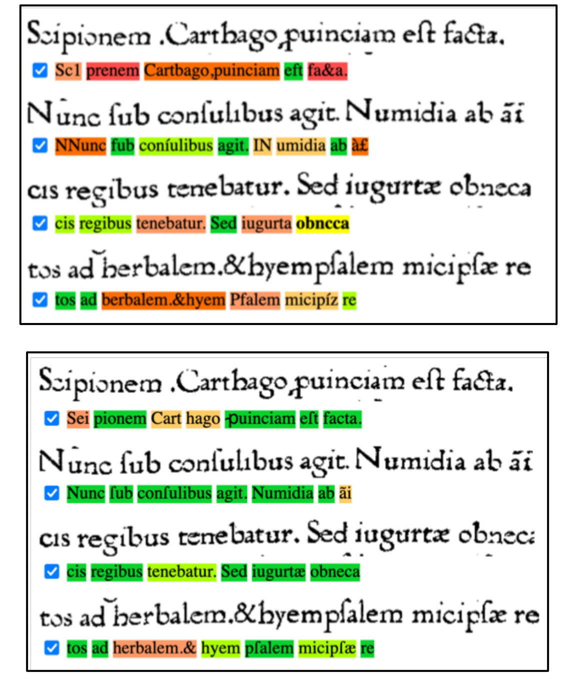

☑ Sc1 prenem Cartbago,puinciam eft fa&a.
☑ NNunc fub confulibus agit. IN umidia ab a£
☑ cis regibus tenebatur. Sed iugurta obncca
☑ tos ad berbalem.&hyem Pfalem micipfz re
☑ Sei pionem Cart hago puinciam eft facta.
☑ Nunc fub confulibus agit. Numidia ab ãi
☑ cis regibus tenebatur. Sed iugurtæ obneca
☑ tos ad herbalem.& hyem pfalem micipfæ re

Figure 3: Word confidence heatmap for Modern Latin model (top) and Historic font model (bottom)

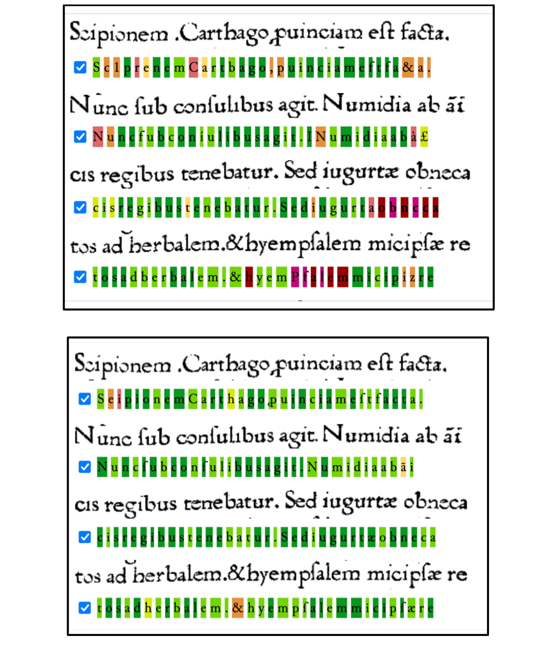


Figure 4: Character confidence heatmaps for Modern Latin model (top) and Historic font model (bottom)

confidence band – and neither one scores consistently higher than the other. Where we might have expected the Latin model to fare better than the English model, the confidence graphs do not hint at a significant nor even a consistent difference. Compared to the output of these two modern models, a Tesseract model trained on historic fonts from books printed in Latin, French and English, shows consistently higher confidence scores, no matter the lack of language focus.[4] This suggests that in this case, the historic font as a parameter in the training material has more impact on the OCR model's word confidence scoring than the focus on language.

In order to test this assumption in more detail, the word confidence scores for the modern Latin model and the historic font model were visualized next to the corresponding text lines using confidence heatmaps: in percentage blocks of 10, descending from 100%, word confidence scores were highlighted in colours shifting from green to red, the former signalling higher and the latter lower scoring (see Figure 3)

Comparing the two heatmaps, the historic font OCR model fares a lot better than the modern Latin OCR model. Yet the isolated problem areas in the historic OCR model map also put the low scores of the modern Latin model into context. Both models seem to struggle with unusual spacing typical for historic printed material, although in different ways. The modern Latin OCR model, moreover, assigns a low confidence score to a lot more words – most, though not all, are incorrectly transcribed or truncated.

In the next step, heatmaps were created to visualize character confidences (see Figure 4). As explained above, the character confidence output from Tesseract's LSTM model is normalized and moves within a smaller band than word confidence. The colour gradation changed per percentage point, from 99 downwards. When looking at the output from the two models on character confidence heatmaps, we can see more clearly where whole words are scoring low on account of single characters, and where whole groupings of letters are affected – and conversely, where unusual spacing rather than low-confidence characters lead to a low word confidence score. The character confidence heatmap for the Historic font model, for example, shows that all letters in "Carthago" are transcribed at relatively high confidence; the low

[4] The last page where the modern Latin model scores highest seems to contradict this; however, the actual page image does not contain any text and can be ignored in this case.

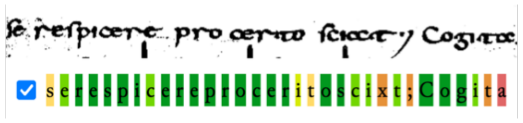

☑ s e r e s p i c e r e p r o c e r i t o s c i x t ; C o g i t a

Figure 5: Confidence heatmap for Tesseract model trained on Carolingian minuscule run over Insular Minuscule manuscript

(correct transcription: "se respicere pro certo sciat; Cogita)

score of the split word presumably stems from the irregular spacing, which divides one legitimate word into two non-words. The character heatmap for the Modern Latin model shows different patterns: in some words, single characters are to blame for a whole word's score, elsewhere whole sequences of letters are affected, such as in the last two lines. In these latter cases, presumably, not only the single letters but also their surrounding characters affect the single character confidence scoring.

The comparison of historic and modern printed font may not yield many surprises but serves as an example for what kind of issues the confidence visualizations can help to surface. Since the introduction of neural networks and LSTM to OCR and HTR technology also opened the door to processing more challenging materials, such as medieval manuscripts, it is conceivable that the same technique can be used for examining differences between different medieval script types. A bespoke model trained on one script type might reliably fail to transcribe unfamiliar ligatures or characters in a test manuscript, flagging up such phenomena with low confidence values (see Figure 5 – the model here struggles with the unfamiliar letter shapes for "a" and "t", untypical for Carolingian minuscule).

What these heatmaps also communicate, however, is that confidence scoring must be taken with a grain of salt: a low confidence score does not always mean that its cause is meaningful for the discussion, nor can we expect to securely identify the cause behind each scoring. Most poignantly, to understand whether an irregularity spotted is a systemic or an anecdotal occurrence, it is necessary to corroborate these findings with more data and evaluate them in the context of the entire document or sample. The next section is presenting a concrete example to showcase how to systematize such an approach with the help of statistical evidence.

2.3 Confidence heatmaps for scribal hands

Palaeographic analysis of medieval and Renaissance manuscripts deals with utterly human material – and to date relies on utterly human expertise. Often, a judgment call is made on account of intuition more than objective grounds. This is not only due to the blurry boundaries between script types and hands but perhaps also owed to the fact that differences between scripts are often difficult to describe or classify objectively. In the analysis of script types or manuscripts hands, visible differences between specific letter forms taken from different exponents can provide means for an objective comparison – except it is not always obvious where to start the search or how to weigh such discoveries.

The above experiment aimed to show how variegation in confidence levels can highlight pages, words, or letters outside the comfort zone of an OCR model. The same mechanism is applied in the following scenario by running a model trained on one manuscript over other exponents. A word-confidence graph maps the general affinity between the chosen manuscripts. Character confidence heatmaps were then used to identify low-scoring letters. Two metrics are used to then evaluate the relevance of these findings: the average confidence measured for all transcriptions of this letter (including scores for correct and incorrect transcriptions); and the overall error rate. These metrics are compared across all manuscripts under review. The average confidence and error rates from the manuscript the model is trained for serve as a baseline against which the results from the other exponents are compared. Said baseline helps to understand whether the reactions of the model cohere with its "comfort zone" or whether they signal a divergence from the baseline.

The process previously described showcased of models trained on a large number of text lines belonging to the same language or script group; the experiment here starts with training a bespoke model for just a single manuscript to then run it over test pages from the same and other manuscripts. Usually, Tesseract models are trained on hundreds of thousands of lines. Granted the small scope (and specific aim) of this experiment, the training of a bespoke model for a manuscript hand was performed with comparatively little material – hundreds, not thousands of lines.

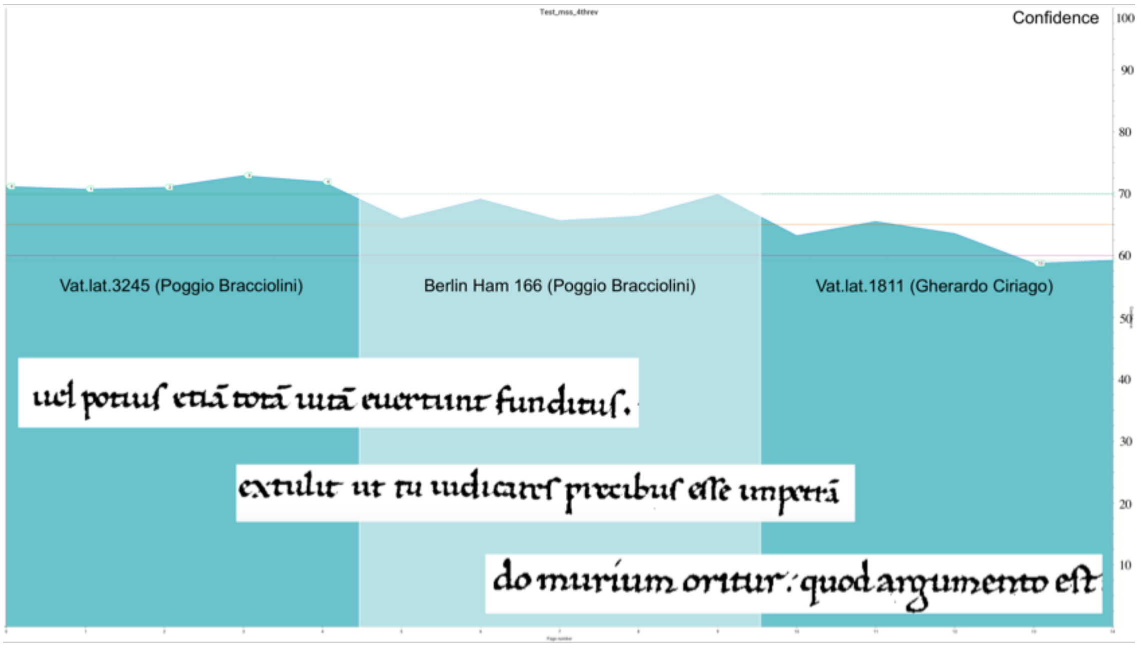


Figure 6: Word confidence graph for Vat.lat.3245, Berlin Ham 166 and Vat.lat.1811 processed with Tesseract model trained on Vat.lat.3235

For this experiment, the hand of Poggio Bracciolini (1380-1459), eminent Humanist and famous scribe, was compared to that of a follower – as also his own. Bracciolini, a trained notary and successful scribe, was the instrumental driver behind the development of what came to be called Humanist Minuscule (de la Mare 1977). Bracciolini is not only a rare example for the deliberate development of an idiosyncratic hand; it is equally uncommon that the evolution of a single hand can be traced with a number of surviving exponents (De Robertis 2017).

Whilst developing his style, Bracciolini also trained (and inspired) imitators. A limited number of manuscripts signed or authenticated by documentary evidence can be securely ascribed to Bracciolini. Based on visual comparison with these manuscripts, numerous others have been identified – and disputed in return (de la Mare 1973, Caldelli 2006). When trying to authenticate manuscripts putatively ascribed to Bracciolini, the palaeographic challenge is a war on two fronts: the first challenge is to identify differences or affinities between two manuscripts; in the second instance, the palaeographer must decide whether these findings signal the same hand or the penmanship of another. The OCR-based methodology cannot provide a secure answer – but as is to be shown, it can be used to identify samples for discussion.

The experiment used a model trained on lines taken from a manuscript identified as an autograph by Bracciolini, Vat.lat.3245 (785 lines for training, 87 lines for evaluation). The resulting OCR model was run over 5 pages each from Vat.lat.3245, Ms. Vat.lat.1811, a manuscript written by his close follower, Gherardo del Ciriago (1412-1472), and another manuscript ascribed to Bracciolini (Berlin Ms Hamilton 166).

The word confidence graph – unsurprisingly, perhaps – suggest that the model generally processed the other Bracciolini autograph at greater confidence levels than the hand of Gherardo del Ciriago (see Figure 6). The gap between the confidence scores for Vat.lat.3245 and Ms Ham 166, however, suggest that the hands, even though belonging to the same person, do differ somehow – perhaps a consequence of them being copied at different stages in Bracciolini's life: Ms Ham 166 was authored in 1408; Ms Vat.lat.3245 is dated to 1410-1415 (de la Mare 1973).

As with the example running OCR on historic printed text, heatmaps were used to understand the confidence scores in more detail. Concretely, in this case, the heatmap provided a first point of contact to help identify letters of interest. In the second step, a closer analysis of the confidence scores for a particular letter was analysed across the whole sample.

A single character transcribed at low confidence would not yield credible data to support an analysis but looking at all transcriptions of the same letter across the test set, i.e. from a variety of contexts, can give a more balanced perspective on whether one or several low confidence ratings signal anecdotal or systematic failure. In a next step, therefore, the confidence values were analysed for all instances of "suspicious" letters across all manuscripts to understand whether we are looking at a meaningful difference or not.

The analytical framework builds on the assumption that the confidence values from Vat.lat.3245 provide the baseline to compare the ratings from the other manuscripts to. In addition to the overall confidence ratings (which included scores for correct and incorrect transcriptions) the error rate is calculated.

Using the heatmaps, following letters were singled out for analysis: "ct", for low scores in Vat.lat.1811; "h", for low scorings in Vat.lat.1811; and "ae" for low scorings in Ham 166. The confidence values and error rates were then collected from all pages in the test set.

In the case of "ct", for Vat.lat.3245 and Ham 166, average confidence and error rate were almost on par. The numbers for Vat.lat.1811, however, differed drastically, particularly the error rate.[5] When comparing samples from the ct ligature across all manuscripts, in fact, the difference is not only consistent but easily visible: the "c" is touching the middle stroke of the "t" (See Table 1).

The case of the letter "h" is less straightforward (see Table 2). Whilst the average confidence level for Ham 166 is not too far from Vat.lat.3245, the error rate is significantly higher; it is also puzzling that two thirds of the errors were transcribed to "b". Overall, the statistics are not definitive enough to support a divergence between Ham 166 and Vat.lat. 3245, nor did the samples surface a regular, visible difference. For Vat.lat.1811, however, the error rate was over 55%. Looking at the manuscript itself, the letter shape regularly differs from the samples found in the other two manuscripts: the belly tends to be wider, and the initiating stroke more horizontal than for the other manuscripts; the final stroke regularly reaches below the baseline.

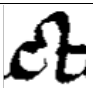
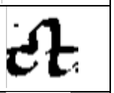
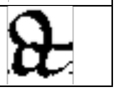
ct ligature	Avg conf	error rate	sample
Vat.lat.3245 (10 total)	98	20%	
Ham 166 (22 total)	97.7	22%	
Vat.lat.1811 (27 total)	95.0	85%	

Table 1: Statistics for "ct" ligature

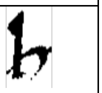
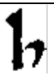
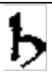
h	Avg conf	error rate	sample
Vat.lat.3245	97.8	7.8% (b: 3.9%)	
Ham 166	97.3	29.2% (b: 18.9%)	
Vat.lat.1811	96.2	55.2% (b: 11.9%)	

Table 2: Statistics for "h"

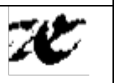
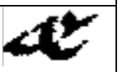
ae	Avg conf	error rate	sample
Vat.lat.3245 (1 total)	97.3	100%	
Ham 166 (24 total)	94.4	100%	
Vat.lat.1811	--	--	--

Table 3: Statistics for "ae" ligature

The "ae" ligature, meanwhile, presented a quite different situation (see Table 3). Initially chosen for low confidence and bad transcription in Ham 166, it turns out that the ligature appears not at all in Vat.lat.1811, and only once in Vat.lat.3245 – and is badly transcribed here, too. Granted "ae" is

[5] The ct ligature is usually transcribed by two letters, so the average confidence rating for both was used in the calculation.

included in the registered set of characters permitted in the Tesseract training, it would not be categorically excluded from recognition. The consistent failure to correctly recognize the glyph thus suggests it is scarce if not absent in the training material to start with – hence the low scores in its recognition. As it is, the use of "ae" or *e caudata* was not obligatory in either manuscript – these were orthographic novelties introduced by Humanist circles in Florence in the 14th and 15th century that aimed to replace the simple "e" common from Medieval times. The relative frequency of this glyph in Ms Ham 166 suggests that Bracciolini chose to deliberately employ it in Ms Ham 166, but mostly reverts to simple "e" in Vat.lat.3245.

Above examples are not exhaustive but give an idea of the kind of material one might identify with the help of OCR confidence scores. Neither graphs nor heatmaps deliver very clean nor comprehensive evidence; they require human scrutiny and interpretation to yield up useful information. In that sense, the examples above do not intend to provide a clear-cut interpretation of the relationship between the three manuscripts. Rather, the intent is to showcase how different modes of visualizing confidence scores from OCR processing can aid the quest for material to feed into palaeographic analysis. How this evidence is ultimately to be weighed is left to the expert; however, the hope is that OCR confidence scores can serve as a heuristic tool to speed up the task in the first place.

3 Conclusion

In summary, this article presented an approach to re-purposing OCR technology for identifying peculiarities in historic scripts or differences in scribal hands. The argument aims to show that even though many standard OCR models, in this case Tesseract, are overtly trained to focus on the recognition of specific languages in print, the sensitivity of OCR models to differences in fonts can be exploited to highlight differences in script or scribal hand. This is done with the help of the so-called confidence score, which signals the certainty with which the OCR engine assumes the output to be correct. The argument above is outlining several methods of visualizing the confidence score and how this can aid palaeographic analysis. The method is emphatically not intended to classify scripts or to authenticate hands. The experiments merely test confidence scores for their heuristic potential in identifying differences between script types and hands.

There are some downsides to this approach; firstly, the confidence score is a blurry metric that can be influenced by many factors, not all of which are relevant to a palaeographic discussion (for example ink bleed-through, speckles or skewed pages). Which factor is chiefly to blame for a low confidence score is not necessarily clear. Gathering scores from every single exponent, by default from a variety of contexts, however, can help to gain a balanced perspective in that regard. Expectations should also be tempered with a view to comprehensiveness – OCR confidence scores cannot be expected to highlight *all* differences in a font or hand. In that sense, the method as sketched presents a means to break the ice.

From a practical perspective, creating ground truth to train bespoke models is time-intensive and low-volume models such as the one used here are not necessarily reliable. Within the small scope of this investigation, such a model might have been sufficient for proving a concept, but a better model trained on more ground truth or endorsed by more advanced technology might be needed for a more thorough analysis. With the steady advance of OCR technology and more and more sophisticated attempts to create models for low-volume scripts, it stands to hope that there will be new solutions to the latter issue before too long.

Lastly, one might argue that these experiments merely help to surface phenomena that are visible to the naked eye anyway. Without providing analytical value in itself, the method leaves the ultimate interpretation of these discoveries to the palaeographer's expertise. It should not be forgotten, however, that palaeographic analysis is a painstaking process, and the identification of such differences is extremely time-consuming when done by hand and from scratch. The real value of this method, therefore, is to direct said naked eye to the phenomena in the first place, that is, to speed up the discovery process – and at the best of times help palaeographers discover elements they did not know they were looking for.

References

Breuel, Thomas M. 2008. 'The OCRopus Open Source OCR System'. Proceedings Volume 6815, Document Recognition and Retrieval XV (68150F). https://doi.org/10.1117/12.783598.

Caldelli, Elisabetta. 2006. Copisti a Roma Nel Quattrocento. Roma: Viella.

Camps, Jean-Baptiste, and Florian Cafiero. 2014. "Genealogical Variant Locations and Simplified Stemma: A Test Case." In Analysis of Ancient and Medieval Texts and Manuscripts: Digital Approaches, edited by Tara Andrews and Caroline Macé, 69–94. Turnhout: Brepols.

Castro Correa, Ainoa. 2014. "Palaeography, Computer-Aided Palaeography and Digital Palaeography." In Analysis of Ancient and Medieval Texts and Manuscripts: Digital Approaches, edited by Tara Andrews and Caroline Macé, 69–94. Turnhout: Brepols.

Christlein, Vincent. 2018. "Handwriting Analysis with Focus on Writer Identification and Writer Retrieval." PhD of Engineering, Erlangen-Nürnberg: Friedrich-Alexander-Universität.

Cilia, Nicole Dalia, Claudio de Stefano, Francesco Fontanella, Claudio Marocco, Mario Molinara, and Alessandra Scotto di Freca. 2019. "A Two-Step System Based on Deep Transfer Learning for Writer Identification in Medieval Books." In CAIP 2019, LNCS 11679, edited by M. Vento and G. Percannella, 305–16. Springer Nature. https://doi.org/10.1007/978-3-030-29891-3_27.

"ICDAR2017 Competition on the Classification of Medieval Handwritings in Latin Script." 2017. Classification of Medieval Handwritings in Latin Script (blog). 2017. https://clamm.irht.cnrs.fr/icdar-2017. [accessed 30 September 2023]

Kestemont, Mike, Vincent Christlein, and Dominique Stutzmann. 2017. "Artificial Palaeography: Computational Approaches to Identifying Script Types in Medieval Manuscripts." *Speculum* 92 (1).

de la Mare, Albinia. 1973. The Handwriting of Italian Humanists. Oxford: Oxford University Press.

de la Mare, Albinia. 1977. 'Humanistic Script: The First Ten Years'. In Das Verhältnis der Humanisten zum Buch, edited by Fritz Krafft and Dieter Wuttke, 89–110. Boppard: Harald Boldt.

Perveen, Shaheen. n.d. 'TESSERACT'. HackMD (blog). https://hackmd.io/@rDplrV2BTM-mnyMNpr9TfQ/Hy4ccns2I. [accessed 30 September 2023]

de Robertis, Teresa. 2017. "Scritture Umanistiche Elementari (e Altro)." Scrineum Rivista 14. http://dx.doi.org/10.13128/Scrineum-21994.

Smith, Ray. 2007. 'An Overview of the Tesseract OCR Engine'. In Document Analysis and Recognition, 2007. ICDAR 2007., 2:629–33. IEEE.

Smith, Ray. 2016. 'Tesseract Blends Old and New OCR Technology'. Tutorial presented at the Document Analysis Systems, Santorini.

Tesseract, https://github.com/tesseract-ocr [accessed 30 September 2023]

'Tesseract Documentation FAQ'. n.d. https://tesseract-ocr.github.io/tessdoc/tess3/FAQ-Old.html. [Accessed 16 November 2023.]

Ul-Hasan, Adnan, Faisal Shafait, and Thomas Breuel. 2013. 'High-Performance OCR for Printed English and Fraktur Using LSTM Networks'. In Proceedings of the International Conference on Document Analysis and Recognition, ICDAR. 10.1109/ICDAR.2013.140.

A Appendix A: OCR models used

rescribev9_fast.traineddata via Rescribe Desktop tool: https://rescribe.xyz/rescribe/ [accessed 30 September 2023]

lat.traineddata and eng.traineddata: https://github.com/tesseract-ocr/tessdata [accessed 30 September 2023]

carolinemsv1_fast.traineddata: https://rescribe.xyz/rescribe/trainings.html [accessed 16 November 2023]

B Appendix B: Test Books and manuscripts

Boccaccio, Giovanni. 1585. *Il Decameron*. Giunti: Venice.

Festus, Rufius. 1472. *Breviarum rerum gestarum populi* Romani: Venice. https://digitale-sammlungen.de/en/view/bsb00006378?page=,1 [accessed 30 September 2023]

Rome, Vatican Library, Ms Vat.lat.3245

Rome, Vatican Library, Ms Vat.lat.1811

Berlin, Ms Hamilton 166

Einsiedeln, Stiftsbibliothek, Codex 281(886): Ascetica; Glossa psalmorum; Poenitentiale (https://www.e-codices.unifr.ch/en/list/one/sbe/0281)

Emil.RuleZ! – An exploratory pilot study of handling a real-life longitudinal email archive

Balázs Indig[1,2,5], **Luca Horváth**[2,5], **Dorottya Henrietta Szemigán**[2,3], **Mihály Nagy**[2,4]
[1]Eötvös Loránd University, Department of Digital Humanities
[2]National Laboratory for Digital Humanities
[3]Eötvös Loránd University, Doctoral School of Literary Studies, Comparative Literature Doctoral Program
[4]Eötvös Loránd University, Atelier Department of Interdisciplinary History
Muzeum krt. 6-8., H-1088, Budapest, Hungary
[5]Eötvös Loránd University, Doctoral School of Informatics
Pázmány Péter stny. 1/C, H-1117, Budapest, Hungary
{indig.balazs,horvath.luca,szemigan.dorottya,nagy.mihaly}@btk.elte.hu

Abstract

An entire generation that predominantly used email for official communication throughout their lives is about to leave behind a significant amount of preservable digital heritage. Memory institutions in the USA (e.g. Internet Archive, Stanford University Library) recognised this endeavor of preservation early on, therefore, available solutions are focused on English language public archives, neglecting the problem of different languages with different encodings in a single archive and the heterogeneity of standards that have changed considerably since their first form in the 1970s. Since online services enable the convenient creation of email archives in MBOX format it is important to evaluate how existing tools handle non-homogeneous longitudinal archives containing diverse states of email standards, as opposed to often archived monolingual public mailing lists, and how such data can be made ready for research. We use distant reading methods on a real-life archive, the legacy of a deceased individual containing 11,245 emails from 2010 to 2023 in multiple languages and encodings, and demonstrate how existing available tools can be surpassed. Our goal is to enhance data homogeneity to make it accessible for researchers in a queryable database format. We utilise rule-based methods and GPT-3.5 to extract the cleanest form of our data.

1 Introduction

We live in a time when many people's email correspondence is preserved as a digital legacy. As these people have mostly used email for official communication throughout their lives (because of their habits) it is possible to look back over the majority of their electronic written communication (Jaillant, 2019). However, such digital email legacy raises a number of moral and legal questions. For example, data protection and privacy legislation, such as the General Data Protection Regulation (GDPR) (European Union, 2016), demands a challenging compliance process and encourages institutions not to implement long-term preservation for their own safety, which may become a relevant issue in the future. Furthermore, in most cases, there is not necessarily a complete separation between private and corporate emails (Cocciolo, 2016; Srinivasan and Baone, 2008). While from the technical perspective, email has undergone many changes since its inception and has become heterogeneous in terms of standards and implementation (Partridge, 2008), which makes longitudinal archives difficult to process, analyse and aggregate.

While different tools can be used to create email archives in different formats (Digital Preservation Coalition and Prom, 2019) in our case it is assumed that the mail archive is already available as an MBOX file (e.g. Google Takeout or similar services). We got legal authorisation to use (without publishing partially or fully) a real-life email archive of a deceased public figure's correspondence (Hungarian, Romanian, English) for our pilot study to uncover and solve possible technical difficulties. Our aim was to create a methodology that could successfully process a real-life MBOX file that contains a longitudinal correspondence and produce an output that could be searched, visualized, and analyzed by researchers interested in the author's official communication.

2 Evaluation of the Available Tools

Several tools exist for processing MBOX archives and all of them are built by adopting alternative ap-

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 172–178
December 1-3, 2023 ©2023 Association for Computational Linguistics

proaches with specific perspectives in mind. Since these approaches are very diverse, each tool has its own strengths and weaknesses which should not be ignored when pursuing our goal (Carlson, 2020). In this section, we introduce a selection of the tools we evaluated (divided into two classes) before we decided to write our own.

For each test we used two MBOX files. One is an artificially created demo MBOX file (Willison, 2022b) containing only two emails. It lacks misspelled emails, notifications, circulars, returned error messages from mailer-daemons (i.e. mail delivery systems), etc. The other is the aforementioned real-life MBOX file that we want to process covering more than one decade and therefore containing many of the errors mentioned above.

All examined tools more or less handle the important metadata headers (FROM, TO, CC, BCC, SUBJECT, DATE, etc.) but most of them fail when it comes to decoding the textual data that particular email bodies contain. We can classify these tools into two main classes: a) those often abandoned and only half done, and b) the complex monolith solutions of institutions that are hard to use without appropriate expertise. We will use this classification in the rest of this chapter.

The first group of programs (Vestal, 2018; Willison, 2022a; Sharma and Bhattacharya, 2023; Mineev, 2021; Juopperi, 2016) promises to process MBOX files and insert the output into various data formats (e.g. CSV, JSON, SQL) for further processing. For the artificial data set, this group of tools worked well but with our non-artificial test data, they failed to produce usable output (i.e. contained raw undecoded string fragments e.g. in *Quoted-Printable* or *Base64* form) if any. We examined the root cause of the errors and it turned out that these programs apply false assumptions and are beyond repair. In general, if one wants to process a very complex data set, their use is not recommended without technical expertise, and we decided it was better to start with a clean slate.

For people without technical skills, viewing an MBOX file in *Mozilla Thunderbird* or *MBOX Viewer* is a great opportunity to interact with the data (i.e. read it), as these tools are halfway between the two mentioned groups: they can handle non-artificial data and do not require expertise (neither technical nor archiving) to operate. However, we found that the export functionality of MBOX Viewer is half broken: it can produce a CSV file for our non-artificial data set, but we could not properly load it with *MS Excel* or *LibreOffice Calc* probably because the garbled delimiters, limiting further deeper analysis. We assumed that it did not escape characters with syntactical meaning in tabular format, breaking the data structure.

In the second group, we tested two well-established solutions that promise more than extracting MBOX to common data formats. *Mailbagit* (University at Albany, 2023b) can import various kinds of email archives and convert them to *MAILBAG* format (University at Albany, 2023a) while exporting the data to other formats (TXT, WARC, PDF, etc.). For our non-artificial data set it yielded a lot of error messages, but could produce a good CSV file with the mandatory metadata and individual files for the payload of each message. When necessary it uses automatic character encoding detection but detects absurd encodings (e.g. Turkish code page), and there is no way to correct such mistakes manually due to its complexity.

ePADD (Stanford University's Special Collections, 2021; Schneider et al., 2019) is developed by the Stanford University Library primarily for English email archives. It offers various features for importing data and allows you to choose how and from where to import it. The same errors were identified as in Mailbagit, but it offered no possibilities for manual repair. It is a comprehensive solution when it comes to email archiving and is the most advanced tool we could find for those who possess archiving expertise but lack technical skills.

In the long run, the bugs may be fixed in some of the aforementioned programs by their maintainers. However, as they did not fit our primary goals, we decided to implement our own lightweight solution for this specific archive which can be easily extended with little technical skills if an error occurs while processing other archives.

3 Our method

To accommodate storing a sequence of multiple emails, the format of raw MBOX data contains encoded fragments, i.e. one-byte long ASCII characters with no syntactical meaning chunked into 80-byte long lines, which allows easy handling. Non-ASCII byte sequences are encoded by a *binary-to-text encoding* scheme such as *Quoted-Printable* or *Base64* which when decoded returns the original values (i.e. string or binary) of the individual emails. One email record is composed of multiple case-insensitive key-value-style (standard and non-standard) headers and the payload consisting of recursively embedded payload parts (e.g. HTML, text, or binary attachment). To get the character representation of non-ASCII, byte-represented text segments, bytes need to be decoded with the supplied character encoding. In some cases, no or wrong character encoding was specified e.g. binary data erroneously has character encoding. We used the built-in *email* module of Python which contains utility functions for most of the aforementioned steps. Proceeding carefully from the headers to the payload we realised they posed different challenges, therefore in the following, we discuss them separately.

3.1 Headers

We gathered the frequency distribution of all headers and their values in a case-insensitive manner. This showed us the irregularities of the data and the non-standard headers which were to be normalised or ignored. We classified headers into three types which we identified by the name of their keys: a) date, b) plain text, and c) address list.

The date values in some cases contained localised human-readable statements on the timezone – sometimes with additional character encoding errors – or contained timezone information in a form that was not handled by Python (-0000 instead of +0000 for UTC time). Fortunately, there is a specific built-in function (*email.utils.parsedate_to_datetime()*) in the email module that handles all but the negative UTC case which we replaced beforehand.

The *email.header.decode_header()* built-in function did the heavy lifting on decoding the abovementioned binary-to-text encoded plain text chunks while keeping the data in bytes form with their specified character encoding because only binary-to-text encoding is safely decipherable. The returned chunks had to be converted to character strings with our own code using the supplied character encoding and handling possible encoding errors. In some cases, the built-in function did not return any character encoding. This either meant strings were already decoded (i.e. were in string type) or it was left up to us to interpret the remaining bytes-type part (in our case all of which were in ASCII). Finally, string chunks needed to be concatenated to restore the original value.

The address sequences could be uniformly split with the *email.utils.getaddresses()* function to name-address pairs (the format defined in the email standards). However, this function leaves the decoding of the binary-to-text data to the user, therefore, the aforementioned decoding heuristics had to be reused here.

These methods cover the common header types, which are required for average use cases. Our program lets the user include or map non-standard headers at will for special use cases (e.g. thread-id, delivered-to, etc.) as they probably do not need extra decoding steps.

3.2 Payload

The email payload is recursively built from *parts*. Nowadays, most emails' body is in HTML, but have a plain-text variant as a separate part which may or may not represent the same textual content as the HTML. Binary parts are also common due to attachments or inline elements (e.g. images) which HTMLs are often augmented with. These components are stored as individual payload parts, but are difficult to distinguish them. As our goal was to extract text only we could ignore attachments and inline binary blobs. However, according to the standards only one of the HTML or plain text content is required – but both are allowed and commonly used side by side –, therefore we decided to keep all textual information and examine them later.

For each payload part, we created a table of values of the available features (which can be extracted by the built-in functions) to define the behaviour of our program by inspecting the groups of values. The used features are listed in Table 1.

Name	**Value**
filename	str/None
is_multipart	True/False
content_type	MIME-type/None
payload	bytes+encoding/None
content_charset	str/None
content_disposition	str/None
has_parts	True/False

Table 1: Features used to classify payload parts

We found significant connections between the features. Some were expected[1], but others were not. For example, *content_disposition* turned out to be unusable as it had inconsistent values, therefore,

[1] *is_multipart* and *has_parts* had the same values: when they were True *payload* was None.

we used the *filename* instead. Besides that, *content_type* were missing in some cases, so we had to utilise *libmagic* to detect one. We broke down the data set into the following classes (Table 2.):

Class name	**Action**
multipart	recurse on subparts
filename is not None	attachment, ignore
content_charset is None and not text MIME-type	inline image, ignore
content_charset is None and text MIME-type	text w/o encoding
content_charset is not None and strange MIME-type	textual data (CSV, iCal, etc.)
not multipart and no filename and content_charset is not None and text MIME-type	proper text

Table 2: Payload class-action pairs

4 Using GPT for Curation

The remaining problems turned out to be the result of non-standard behaviour, which was easier to solve with a solution more capable of handling semi-structured data, therefore we chose to use *GPT-3.5* (Ouyang et al., 2022) (to facilitate affordable reproducibility) adopting a few-shot methodology (see Appendix A.1.). Upon replying to, or forwarding an email the old text is separated by the main metadata of the email (in string form formatted in a non-standard, language-specific way) from the new text which is usually written at the top of the email body, however, in some cases at the discretion of the user, the reply is inlined (inserted between the lines of the old email). This process often results in concatenated email-body texts requiring separation. To solve these issues, we identified the following tasks: a) decoding the use of non-standard string form metadata inside the email body, and b) separating concatenated email-body texts. Furthermore, we also used GPT-3.5 to fix the remaining edge-case encoding errors that were caused by "lossy" decoding (e.g. replacement characters, omitting faulty byte sequences, incorrectly decoded characters).

GPT-3.5 performed best on metadata extraction, where it was able to extract FROM and TO addresses, dates, and SUBJECT strings even with simple prompts. Handling concatenated email-body texts, however, proved to be a more difficult task, therefore, for testing their separation accuracy, we chose 100 examples with varying numbers of previous email text recursively included in the payload. 77 were successfully separated, failing mostly on inputs with more complex text structures that did not contain proper helper annotation (e.g. ">" used for indenting previous email text lines) (see Figure 3. in Appendix A.1.).

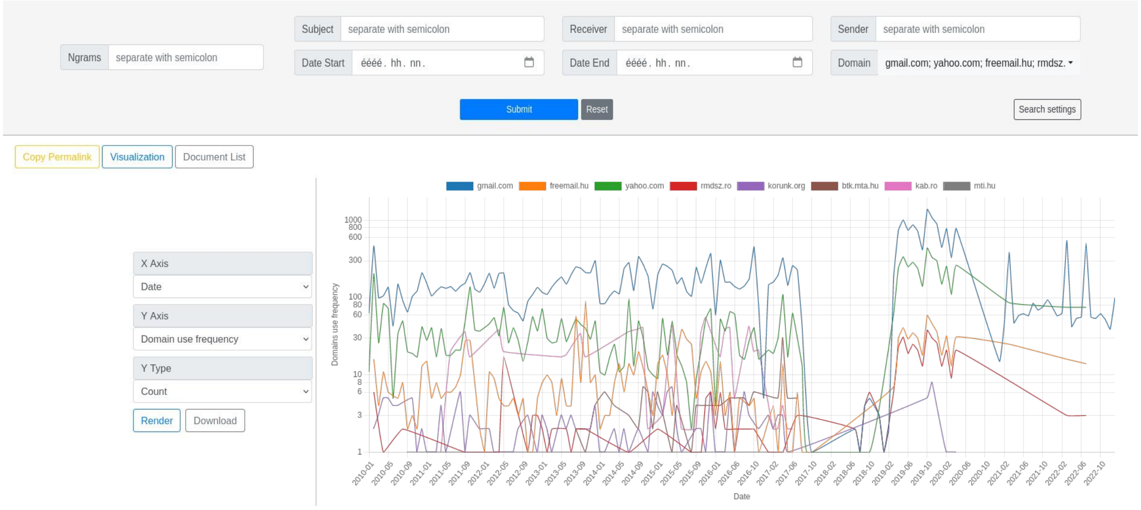
We found that many of the erroneously encoded emails that were left had the same problem: an automatic mechanism (antivirus) had pasted a footnote to the payload, but with wrong encoding, which caused the decoding of the whole message to fail. Although GPT-3.5 successfully handled these cases, we also implemented a rule-based method of splitting the text and applying another encoding to the footnote part, fully eliminating this type of encoding problem. Only a few complex encoding errors remained that neither GPT-3.5 nor a rule-based method could solve, as they were products of several layers of incorrect processing, and the resulting character combinations were indecipherable even for humans. Our experiments show that a rule-based workflow could potentially be expanded by using Large Language Models, if tasks are well compartmentalised and split into separate problem areas.

5 Visualising the Resulting Data

With the data cleaned and normalised to the limit, as a final step to facilitate access to authorised digital humanists who prefer visual representations of data, we loaded it into an off-the-shelf application suitable for n-gram based data exploration (*N-gram Trend Viewer* (Indig et al., 2022)). One example of exploring the metadata-rich text-based corpora – using only metadata that is safe from compromising GDPR – is the frequency of different email providers that the owner of the account interacted with over time (see *Figure 1.*). Naturally, those who have legal access to the data can make more in-depth queries that the system supports.

6 Evaluation and Conclusion

Our experiments with the automatic character encoding recognizer systematically resulted in Turkish code pages, which can be safely ruled out from the set of possibly used languages and code pages, therefore, we opted to manually observe each oc-



Ngrams: separate with semicolon

Subject: separate with semicolon

Receiver: separate with semicolon

Sender: separate with semicolon

Date Start: éééé. hh. nn.

Date End: éééé. hh. nn.

Domain: gmail.com; yahoo.com; freemail.hu; rmdsz.

X Axis: Date

Y Axis: Domain use frequency

Y Type: Count

Figure 1: Usage count of selected anonymized email domains (FROM, TO, CC, and BCC) over time. There was a drop in traffic around 2017 when the subject passed away, however, strangely enough after a period of inactivity the account started interacting with several domains again, most likely due to someone gaining access.

currence and guess the most likely encoding. With the described payload classification only a few (37 from 11,245) email payload parts remained that had trivially erroneous encoding specified, or failed to decode with the specified/suggested encoding. This could be further reduced by using GPT and the string-splitting technique. The existing methods we evaluated found the same number of character encoding errors as our method did, however, the final error rate was worse for the evaluated methods due to the automatic mechanisms and the lack of possibility for correction.

Our method and ePADD both found that many email addresses had different names associated with them, which enables these names to be added as an alias along with a canonical name to a semantic database (e.g. the "Also known as" field in Wikidata (Vrandečić and Krötzsch, 2014)) for later use. ePADD used English word lists with little success to recognise named entities in the text. The lists can be changed but due to the monolithic nature of the program, the clearly not state-of-the-art method cannot.

We conclude our pilot project a success[2], as we recovered most of the errors and created an intuitive WebUI for the MBOX data to help researchers explore the email archive. To open up more possibilities in the future, the conversion of emails to standard TEI XMLs (DeRose, 1999) is an option worth exploring as it could additionally handle the complex philological aspects of inline replies. Trying other MBOX files is also desirable to make our tool more robust and handle more non-standard headers since it was built with extensibility and customisability in mind.

Acknowledgements

We extend our special thanks to András Hajós for coining the name of his band, paying homage to the beauty of the email. We hope to express our appreciation by choosing our suggestive title.

This work is supported by the *DKOP-23 Doctoral Excellence Program* of the *Ministry of Culture and Innovation, Hungary* from the source of the *National Research, Development and Innovation Fund*.

MINISTRY OF CULTURE
AND INNOVATION

References

M. E. Grenander Department of Special Collections & Archives University at Albany. 2023a. Mailbag specification (1.0). https://github.com/UAlbanyArchives/mailbag-specification.

M. E. Grenander Department of Special Collections & Archives University at Albany. 2023b. Mailbagit.

[2] The code is published under GPL 3.0 license at https://github.com/elte-dh/mbox-parser.

https://github.com/UAlbanyArchives/mailbagit. Last accessed on 25/09/2023.

Clare Carlson. 2020. One size does not fit all: Exploring email archiving workflows. School of Information and Library Science (master's thesis).

Anthony Cocciolo. 2016. Email as cultural heritage resource: appraisal solutions from an art museum context. *Records Management Journal*, 26(1):68–82.

Steven DeRose. 1999. XML and the TEI. *Computers and the Humanities*.

Digital Preservation Coalition and Christopher J. Prom. 2019. Preserving email 2nd edition. *DPC Technology Watch Report*, 28.

European Union. 2016. Council regulation (EU) no 679/2016. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

Balázs Indig, Zsófia Sárközi-Lindner, and Mihály Nagy. 2022. Use the metadata, Luke! – an experimental joint metadata search and n-gram trend viewer for personal web archives. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 47–52, Taipei, Taiwan. Association for Computational Linguistics.

Lise Jaillant. 2019. After the digital revolution: working with emails and born-digital records in literary and publishers' archives. *Archives and Manuscripts*, 47(3):285–304.

Jari Juopperi. 2016. E-mail message to JSON converter. https://github.com/jmjj/messages2json. Last accessed on 13/09/2023.

Vsevolod Sebastian Mineev. 2021. Python script to extract emails from an .mbox file. https://github.com/vsevolod-mineev/csv-from-mbox. Last accessed on 13/09/2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Craig Partridge. 2008. The technical development of internet email. *IEEE Annals of the History of Computing*, 30(2):3–29.

Josh Schneider, Chance Adams, Sally DeBauche, Reid Echols, Callum McKean, Jessica Moran, J., and Dorothy Waugh. 2019. Appraising, processing, and providing access to email in contemporary literary archives. *Archives and Manuscripts*, 47(3):305–326.

Prakhar Sharma and Adrita Bhattacharya. 2023. MBOX to JSON. https://github.com/PS1607/mbox-to-json. Last accessed on 13/09/2023.

Arvind Srinivasan and Gaurav Baone. 2008. Classification challenges in email archiving. In *Rough Sets and Current Trends in Computing*, pages 508–519, Berlin, Heidelberg. Springer Berlin Heidelberg.

University Archives Stanford University's Special Collections. 2021. ePADD, email Process Appraise Discover Deliver. Last accessed on 25/09/2023.

Allan James Vestal. 2018. mbox-tools. https://github.com/DallasMorningNews/mbox-tools. Last accessed on 13/09/2023.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Simon Willison. 2022a. mbox-to-sqlite. https://github.com/simonw/mbox-to-sqlite. Last accessed on 13/09/2023.

Simon Willison. 2022b. Sample MBOX. https://github.com/simonw/mbox-to-sqlite/blob/main/tests/enron-sample.mbox. Last accessed on 13/09/2023.

A Appendices

A.1 The Used Prompt for Email Separation

MODEL gpt-turbo-0613
TEMPERATURE 0
SYSTEM You are a [LANGUAGE] email analysis assistant. Your job is to take an email body text and if it contains the text of previously sent emails as an email correspondence, then separate it into individual emails, and finally return the separate emails annotated with ”SEPARATE_EMAIL:”. You also need to remove any characters that are in some cases added to the email body text to annotate forwarded emails, emails sent as replies, or references to the original email. These characters are usually a greater-than character: ”>”. Not all previous correspondences are annotated with a greater-then character. If encounter strings following the format ”[DATE] [NAME] wrote, [EMAIL]:” then leave it in the output as if it were part of the email text. Some input texts may be a single email, others may be a sequence of emails that contain the latest email and other previously sent emails that are replies, forwards, or the original message.
USER-ASSISTANT pairs of example email and example email with separation tags (x 3)
USER {input_email_text}

Table 3: Example prompt details of *OpenAI chat API* requests for separating augmented email payload texts, with a few-shot approach. USER-ASSISTANT email pairs are omitted for privacy reasons. Fourth *USER input_email_text* is to be replaced for each request with the actual email payload text.

Banning of ChatGPT from Educational Spaces: A Reddit Perspective

Nicole Miu Takagi
Waseda University
miu.n.takagi@toki.waseda.jp

Abstract

With the introduction of ChatGPT on November 30, 2022, the online sphere was disrupted seemingly overnight, with its ability to generate human-like text and comprehensively answer questions. It has even been lauded as being able to aid in the editing and generation of code. Some schools and online question-and-answer forums, however, have banned its use. In this paper, we use Reddit data to examine the impact that the banning of the AI tool has had online early in the introduction of the AI, when it began to gain popularity. Our findings indicate that reactions have ranged from skepticism that the ban will work, loss of educational opportunity, to agreement that ChatGPT is not 100 percent accurate in its answers. Our preliminary findings reveal that based on Reddit discussions, it may be postulated that while it may be better to ban it from Question-and-Answer forums, in physical classrooms, ChatGPT may be seen to hinder students from finding their own solutions to problems, and may also provide the opportunity for students to critically view answers provided to them by the chatbot, strengthening their digital literacy and critical thinking skills.

1 Introduction

ChatGPT was released on November 30, 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 (Lubowitz, 2023; Baidoo-Anu and Ansah, 2023). As a large language model trained by OpenAI as an interactive chat module to aid in language-related tasks such as answering questions and even aiding in writing code, it has caused concern to learners and educators that it provides easy access to answers with little understanding of the topic, in addition to it occasionally confidently providing erroneous or biased responses unbeknownst to the inquirer.

In this study, topic modeling will be used to explore online reactions to the NYC ban comparing two different groups: educators and programmers. Succinctly put, reactions have mostly been skeptical of the effectiveness of preventing plagiarism and instead, there have been calls for a review and revision of the education system. Meanwhile, when ChatGPT was banned (Kabir et al., 2023) on the question-and-answer website for programmers, StackOverflow, reactions seemed more agreeable with the ban, emphasizing its generation of incorrect answers.

2 Background

Public sentiment towards ChatGPT has been largely positive (Tlili et al., 2023). However, the reaction from educators has been more mixed (Sullivan et al., 2023; Neumann et al., 2023). Even though it has been less than a year since ChatGPT has been publicly available, plenty of academic literature on its impact on education exists already, both for and against its inclusion as a tool in education. The mixed reactions from educators include some heralding it as a progressive step into the future leveling the playing field by providing equitable learning opportunities (Sullivan et al., 2023), and others fearing its potential to reduce analytical and critical thinking skills while promoting academic misconduct (Grassini, 2023; Yu, 2023).

Some researchers believe that the use of ChatGPT in classrooms will increase engagement and interaction and provide a personalized learning experience, but highlight a need for a strong focus on critical thinking skills (Kasneci et al., 2023). Many also worry about not just the generation of wrong information, but also the augmentation of biases that exist in the training data and the risks to pri-

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 179–194
December 1-3, 2023 ©2023 Association for Computational Linguistics

vacy and security (Baidoo-Anu and Ansah, 2023; Yu, 2023).

Kabir et al. (2023) even pitted StackOverflow and ChatGPT against each other in order to determine which platform provides better answers. They found that despite 52% of ChatGPT answers containing inaccuracies, users preferred ChatGPT's answers due to their comprehensiveness. On the other hand, StackOverflow has often been accused of providing rude and condescending answers to user questions (Calefato et al., 2018; Brooke, 2019) so it is plausible that the fact that ChatGPT does not mock the question is preferable to especially beginner programmers unsure of their own skills despite the risk of providing inaccurate answers.

Recent studies have also leveraged Reddit to collect data on posts discussing the usage of ChatGPT, or attempt to apply the AI as an analytical tool in various fields from mental health care, streaming media, content moderation, and applications of such generative AI (Haman et al., 2023; Feng et al., 2023; Choi et al., 2023; Wickham and Öhman, 2022). While this research preceded the introduction of ChatGPT, researchers have also found that educators widely use Reddit given that the website is able to host a wide variety of topics and spaces, suggesting that its use may be influential for educators (Willet and Carpenter, 2020).

Moreover, in order to distinguish various public sentiments researchers have employed topic analysis. Many featured Latent Dirichlet Allocation (LDA) topic modeling (Li et al., 2023; Melton et al., 2021), one of the most popular topic modeling approaches. Of the many models available, BERTopic (Grootendorst, 2022), known for its embedding approach, recently emerged to be quite competitive against others. Research on Twitter data, has evaluated BERTopic to be effective compared to Top2Vec and LDA (Egger and Yu, 2022). Many studies have already successfully utilized BERTopic to elicit trends and topics from data scraped from the Reddit platform (Sarkar et al., 2022; Liu et al., 2022; Kerkhof, 2023).

3 Data and Method

Three Reddit threads posted early on in the ChatGPT discourse, (Reddit, 2023b),[1] (Reddit, 2023c)[2] and (Reddit, 2023a)[3] were selected for data collection due to them having the most upvotes, the Reddit measure of popularity, 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 are from the thread on bans by NYC public schools, while 856 are from the StackOverflow thread with 6.6k upvotes, and 439 are from the StackOverflow thread with 1.5k upvotes. They were collected through PRAW 7.7.1 via Python Reddit API (Bryce Boe, 2023).

Data were first preprocessed: comments were separated to form a single list containing strings of sentences for each thread respectively and to fully utilize sentence-transformers (Reimers and Gurevych, 2019).

Topic modeling was conducted with BERTopic and sentence transformers with the *all-MiniLM-L6-v2* model to get the most out of the sentence embeddings. Intertopic distance maps and topic word scores were first generated to gain an initial understanding of how topics were clustered and the relative c-TF-IDF scores between and within topics (see appendix). Following this, to determine the relationship between topics, and documents within topics, Figures 7, 9, and 8 were generated.

Finally, to confirm the results, LDA through Gensim was utilized as a baseline. 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

[1] Reddit. 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/

[2] Reddit. Stackoverflow to ban chatgpt generated answers with possibly immediate suspensions of up to 30 days to users without prior notice or warning https://www.reddit.com/r/programming/comments/zhpkk1/stackoverflow_to_ban_chatgpt_generated_answers.

[3] Reddit. Chatgpt ai generated answers banned on stack overflow https://www.reddit.com/r/programming/comments/zd71vl/chatgpt_ai_generated_answers_banned_on_stack .

shown in Figure 10 in the appendix. Topic numbers that formed the highest peak, but still remained less than 100, were selected for each thread; 38 topics for the NYC thread, 80 for the larger StackOverflow thread, and 38 for the smaller StackOverflow thread. In addition to the topic models and intertopic distance maps generated, bigrams, trigrams, and co-occurrence networks were also visualized for each thread utilizing nlplot, an analysis a visualization module dor Python (takapy0210, 2022).

4 Results

4.1 New York City Ban

Interestingly, in the topic models generated by BERTopic, Figure 1 top topics included *the usage of calculators in math, Wikipedia and their source*s, *wrong answers* or *wrong question*s, *phones* and *wifi*, *search engines*, and the *writing of essays*. Details can further be elicited about these topics from Figure 7 in the appendix, where the figure further highlights clusters on circumventing firewalls through the use of Virtual Private Networks (VPN), in addition to being able to observe the math-calculator topic clearly.

Meanwhile, for Gensim's topic model, seemed to pick up on the skepticism against the effectiveness of the ban of ChatGPT from public schools in New York City. Top keywords in the thread, seen in Figure 11a 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 the trigrams contained in the appendix. In the trigram, Figure 11b 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, Figure 14, 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.

4.2 StackOverflow Ban

With regards to the thread that was more upvoted, the thread with 6.6k upvotes, 6, top topics can be seen to be about search engines & OpenAI, ChatGPT & translators, rabbits fitting inside a building, GitHub & copilot & mailing, and travelers & inns & understanding of them. Key term searches of these topics through the corresponding data revealed a debate on the functionalities of ChatGPT. With regards to Topic 4 in Figure 6, users were discussing how ChatGPT acted like an experienced translator, in which it would auto-complete the intent/meaning of the text in a human-like manner, referencing (Reynolds and McDonell, 2021), a study on how ChatGPT outperforms significantly in 0-shot prompts namely in translation tasks.

With Gensim's topic model, top keywords observed in Figure 12a were regarding the Turing test, training set or data of ChatGPT, and wrong answers provided by the chatbot. For the trigram, Figure 12b, however, topics were more varied, though common topics were ["incorrect", "lot", cases"], ["gets", "stuff", "wrong"], and ["scary", "confidently", "incorrect"]. However, there were also topics on ["wrong", "prompts", "usually"], in counter to the aforementioned topics, though this was lower ranked. In the co-occurrence network, Figure 15, central keywords were "like", "people", "ai", "answers", and "correct", similar to the NYC ban thread. Outside the central cluster, no major overarching theme could be observed.

Meanwhile, in the thread with 1.5k upvotes, Figure 5, top topics observed were about questions & StackOverflow, Marvel comics & MidJourney, Google web searches, and Jones & GimmickNG. Topic 1 involved a discussion on whether ChatGPT should be banned on StackOverflow, with many users agreeing with its ban due to various reasons such as the platform being community-driven, to the need for human-generated answers to be used as training data for future iterations of generative AI. Meanwhile, Topic 3 involved a Redditor philosophically questioning the stance of users against AI art, stating that the arguments are quite subjective. Topic 6 involved comparisons between Google, the search engine, and ChatGPT, how both may provide inaccurate information, in addition to users needing to know how to type in their prompts or questions to minimize the inaccuracies. Finally, Topic 7 seemed to be a ChatGPT-generated story that a Redditor posted.

Pivoting to Gensim's topic model, in Figure 13a, top keywords were more varied and contained more random words, such as "genocide" and "communist revolution". Top-ranked terms, however, were still related to ChatGPT as a language model

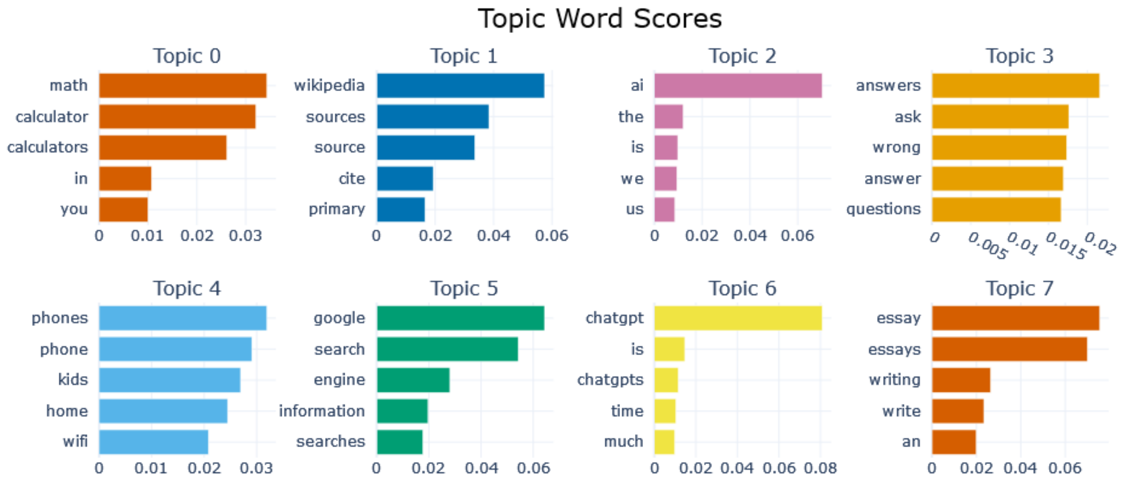


Figure 1: New York Ban topics using BERTopic

and its training data. No major changes in keyword theme could be observed in the trigram, Figure 13b, though topics related to "genocide" were higher ranked compared to the bigram. In the co-occurrence network, Figure 16, similar keywords as the NYC ban and the other StackOverflow thread were observed. Unlike in Figure 13a, the key term "genocide" was not observed to co-occur often. At the bottom of the network, a cluster of terms related to misinformation can be observed, while the left is more about ChatGPT's training model and data.

5 Conclusion and Discussion

5.1 Reddit Responses

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 commonly discussed how calculators were allowed in Mathematics, and the topic on Wikipedia and sources, seemingly discussing how the ban provided a sense of *deja vu* in terms of new tools initially being banned from classrooms. Redditors commenting on the NYC thread were skeptical, raising examples of how students could circumvent the ban. What's more, in the actual thread itself, some have posted methods that ChatGPT has suggested, to demonstrate the ineffectiveness of banning the use of ChatGPT in educational settings. 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, against policing the use of AI, since most were unable to detect work created by ChatGPT. However, the incorporation of the resource may be limited to European-language-speaking or high-resource-language-speaking classrooms due to it being performing not as well as other resources for low-resource languages (Jiao et al., 2023). This, however, does not necessarily mean that AI such as ChatGPT should not be incorporated in low-resource language classrooms. This can instead be taken as an opportunity for students to develop alternative AI, allowing them to practice their skills while also becoming proficient in digital literacy.

StackOverflow's discussion was more technical. Support was shown for the ban due to a lack of accuracy in providing quality code. The term "genocide" is speculated to have been highlighted in the n-grams due to a Redditor posting a song on genocide in the discussion, possibly a Troll[4] posting

[4] a person who provokes others (chiefly on the Internet) for their own personal amusement or to cause disruption (Wikipedia)

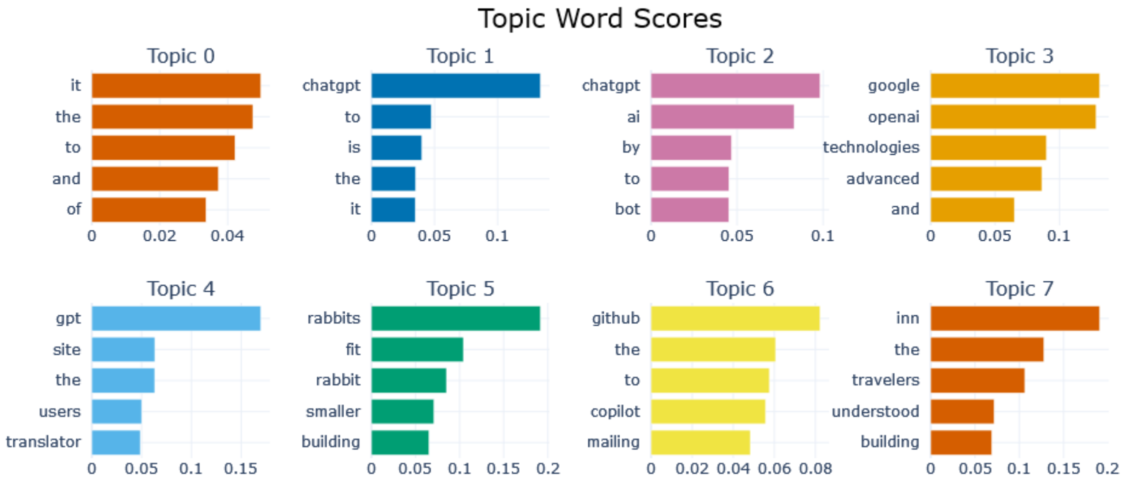


Figure 2: StackOverflow (6k) topics using BERTopic

comments unrelated to the discussion. Meanwhile, "communist revolution" involved a discussion on how AI may be rendered obsolete in true communism. Redditors in this section discussed how the fear of AI taking away jobs in capitalist societies may be solved if a communist revolution occurred. In this manner, this particular thread can especially be seen to highlight how the discussion of AI in educational spaces can diverge into other topics, such as political ideology. In the topic model by BERTopic, "communist revolution" could not be identified while "genocide" was a small cluster in Figure 9, possibly indicating that they are not too relevant to the main topic of discussion. 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 this regard, it may be beneficial for forums on educational platforms to ban the use of AI to allow opportunities for students to learn directly from each other.

5.2 Policy Developments Since the Bans

There have been developments in policy and technology since the bans that occurred in late 2022, early 2023. Following the NYC ban, OpenAI stated their concern for maintaining educational integrity, subsequently releasing an AI checker in late January of the same year, though it was later taken down on July 20, 2023 due to its low rate of accuracy (Kirchner et al., 2023). On May 18, 2023, NYC reversed its ban, a move that was applauded to encourage students and educators to explore new technology (Faguy, 2023). On the other hand, to the knowledge of the authors, the chatbot remains banned on StackOverflow as of November 15 2023.

This divide in policy adoption can be said to be a reflection of the Reddit threads; areas that require expertise prefer human beings, while in those that pursue digital literacy and critical thinking skills, AI may aid the development of skills.

5.3 Conclusion

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.

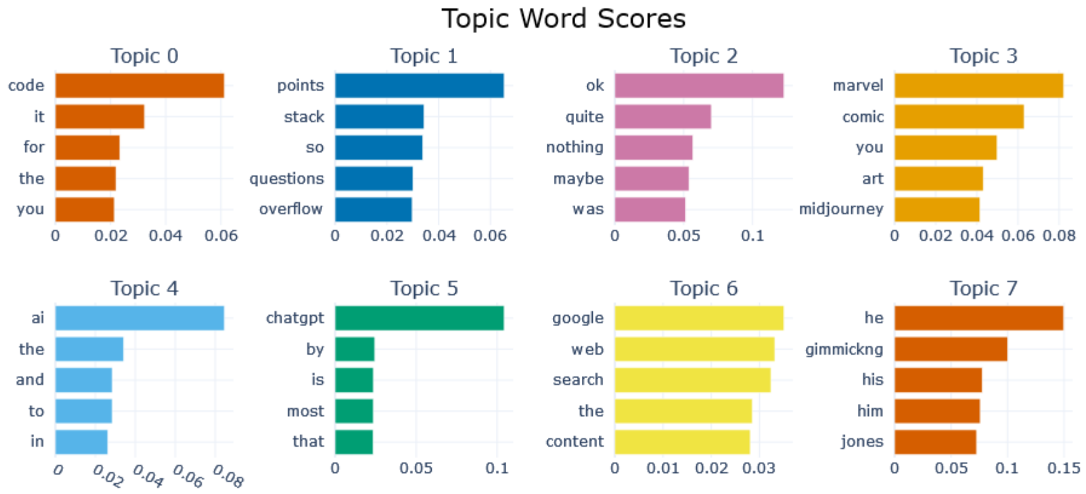


Figure 3: StackOverflow (1k) topics using BERTopic

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.

Limitations

Further research into this topic was limited immediately following the analysis conducted on the data presented in this report due to a strike being conducted by subreddit moderators at the time. Had this data been available, it would have been interesting to see how stances towards ChatGPT have changed over time in educational settings.

Since sentiment analysis is sometimes paired with topic modeling in the research of social media, this is planned to be the next step in the exploration of this data and any extensions to the data collected following the initial data collection.

Ethics Statement

While much care was taken into anonymizing users and ensuring that comments used in the study were not leaked to 3rd parties, it must be noted that publicly available Reddit threads can be easily searched and found. As such, some users may be identifiable, though these ethical concerns and limitations will continue to be considered by the researchers. As such, the researchers will attempt to respond to any concerns raised even after the submission of this paper.

Acknowledgements

The researchers would like to thank Elissa Nakajima Wickham for her support in collecting and analyzing Reddit data.

References

David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *Journal of AI*, 7(1):52–62.

Sian Brooke. 2019. "condescending, rude, assholes": Framing gender and hostility on stack overflow. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 172–180.

Bryce Boe. 2023. PRAW 7.7.1 documentation.

Fabio Calefato, Filippo Lanubile, and Nicole Novielli. 2018. How to ask for technical help? evidence-based guidelines for writing questions on stack overflow. *Information and software technology*, 94:186–207.

Joyjit Chatterjee and Nina Dethlefs. 2023. This new conversational AI model can be your friend, philosopher, and guide ... and even your worst enemy. *Patterns*, 4(1):100676.

Wonchan Choi, Yan Zhang, and Besiki Stvilia. 2023. Exploring Applications and User Experience with Generative AI Tools: A Content Analysis of Reddit Posts on ChatGPT. In *Proceedings of the Association for Information Science and Technology*, volume 60, pages 543–546.

Roman Egger and Joanne Yu. 2022. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in Sociology*, 7.

Ana Faguy. 2023. New York City Public Schools Reverses ChatGPT Ban.

Yunhe Feng, Pradhyumna Poralla, Swagatika Dash, Kaicheng Li, Vrushabh Desai, and Meikang Qiu. 2023. The Impact of ChatGPT on Streaming Media: A Crowdsourced and Data-Driven Analysis using Twitter and Reddit. In *2023 IEEE 9th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, pages 222–227.

Simone Grassini. 2023. Shaping the future of education: exploring the potential and consequences of ai and chatgpt in educational settings. *Education Sciences*, 13(7):692.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

M. Haman, M. Skolník, and T. Subrt. 2023. Leveraging ChatGPT for Human Behavior Assessment: Potential Implications for Mental Health Care. *Annals of Biomedical Engineering*, (51):2362–2364.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is ChatGPT A Good Translator? A Preliminary Study.

Samia Kabir, David N Udo-Imeh, Bonan Kou, and Tianyi Zhang. 2023. Who answers it better? an in-depth analysis of chatgpt and stack overflow answers to software engineering questions. *arXiv preprint arXiv:2308.02312*.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

P. Kerkhof. 2023. Pos0598-hpr what do users of rheumatoid arthritis online forums talk about? applying a deep learning approach to uncover common themes. *Annals of the Rheumatic Diseases*, 82(Suppl 1):570–571.

Jan Hendrik Kirchner, Lama Ahmad, Scott Aaronson, and Jan Leike. 2023. New AI classifier for indicating AI-written text.

Shanghao Li, Zerong Xie, Dickson K. W. Chiu, and Kevin K. W. Ho. 2023. Sentiment analysis and topic modeling regarding online classes on the reddit platform: Educators versus learners. *Applied Sciences*, 13(4).

Yang Liu, Zhiying Yue, and Mohd Anwar. 2022. Monkeypox at-a-glance from google trends and reddit. In *2022 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pages 166–167.

James H Lubowitz. 2023. Chatgpt, an artificial intelligence chatbot, is impacting medical literature. *Arthroscopy*, 39(5):1121–1122.

Chad A. Melton, Olufunto A. Olusanya, Nariman Ammar, and Arash Shaban-Nejad. 2021. Public sentiment analysis and topic modeling regarding covid-19 vaccines on the reddit social media platform: A call to action for strengthening vaccine confidence. *Journal of Infection and Public Health*, 14(10):1505–1512. Special Issue on COVID-19 – Vaccine, Variants and New Waves.

Michael Neumann, Maria Rauschenberger, and Eva-Maria Schön. 2023. "we need to talk about chatgpt": The future of ai and higher education.

Reddit. 2023a. ChatGPT AI Generated Answers Banned On Stack Overflow. [Online; accessed 28-January-2023].

Reddit. 2023b. 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. [Online; accessed 28-January-2023].

Reddit. 2023c. StackOverflow to ban ChatGPT generated answers with possibly immediate suspensions of up to 30 days to users without prior notice or warning. [Online; accessed 28-January-2023].

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm.

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

Shailik Sarkar, Abdulaziz Alhamadani, Lulwah Alkulaib, and Chang-Tien Lu. 2022. Predicting depression and anxiety on reddit: a multi-task learning approach. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 427–435.

Miriam Sullivan, Andrew Kelly, and Paul McLaughlan. 2023. Chatgpt in higher education: Considerations for academic integrity and student learning.

takapy0210. 2022. nlplot: Analysis and visualization module for Natural Language Processing.

Ahmed Tlili, Boulus Shehata, Michael Agyemang Adarkwah, Aras Bozkurt, Daniel T Hickey, Ronghuai Huang, and Brighter Agyemang. 2023. What if the devil is my guardian angel: Chatgpt as a case study of using chatbots in education. *Smart Learning Environments*, 10(1):15.

Elissa Nakajima Wickham and Emily Öhman. 2022. Hate speech, censorship, and freedom of speech: The changing policies of reddit. *Journal of Data Mining & Digital Humanities*.

K. Bret Staudt Willet and Jeffrey P. Carpenter. 2020. Teachers on reddit? exploring contributions and interactions in four teaching-related subreddits. *Journal of Research on Technology in Education*, 52(2):216–233.

Hao Yu. 2023. Reflection on whether chat gpt should be banned by academia from the perspective of education and teaching. *Frontiers in Psychology*, 14:1181712.

A Appendix

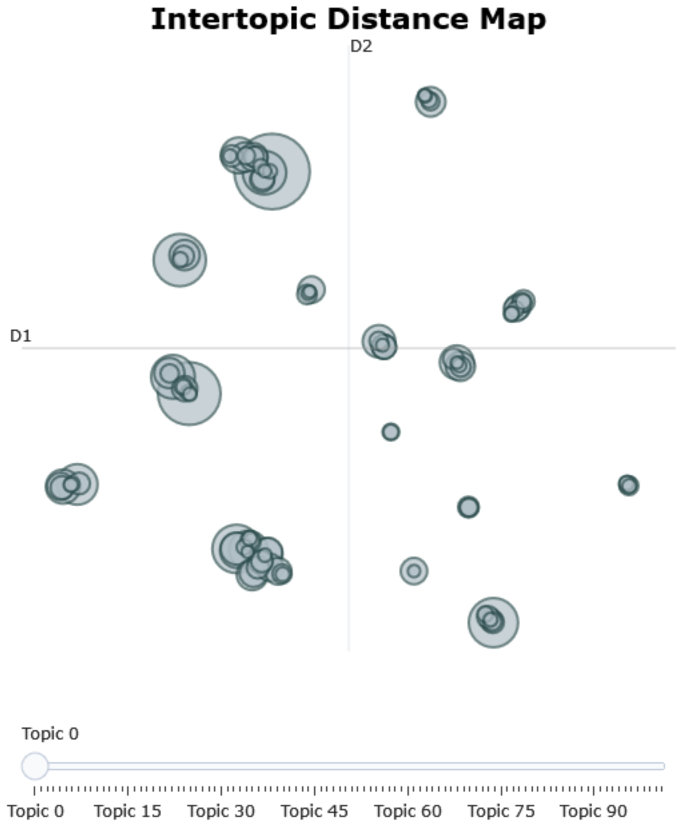


Figure 4: Intertopic Distances in NYC ban Thread

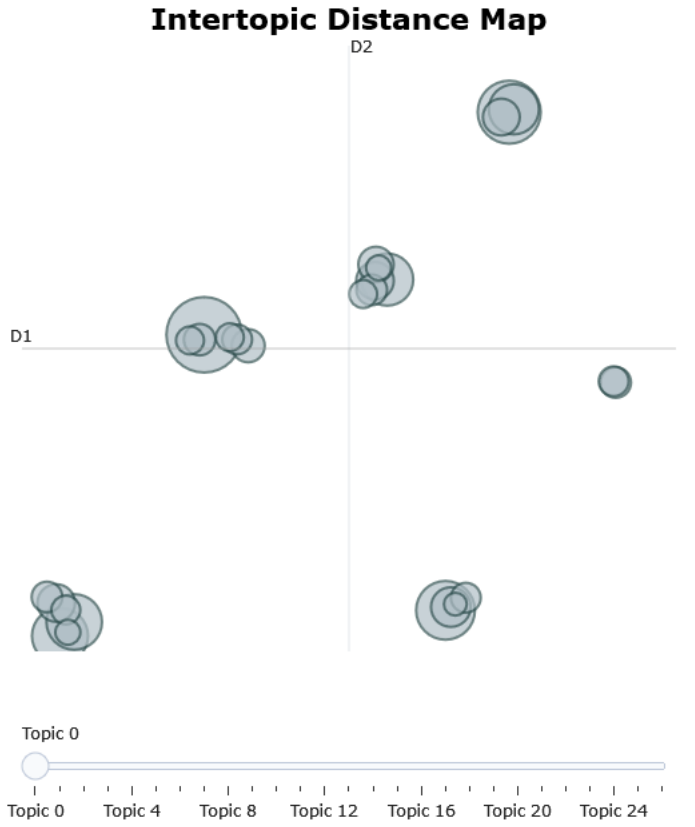


Figure 5: Intertopic Distances in StackOverflow (1k) ban Thread

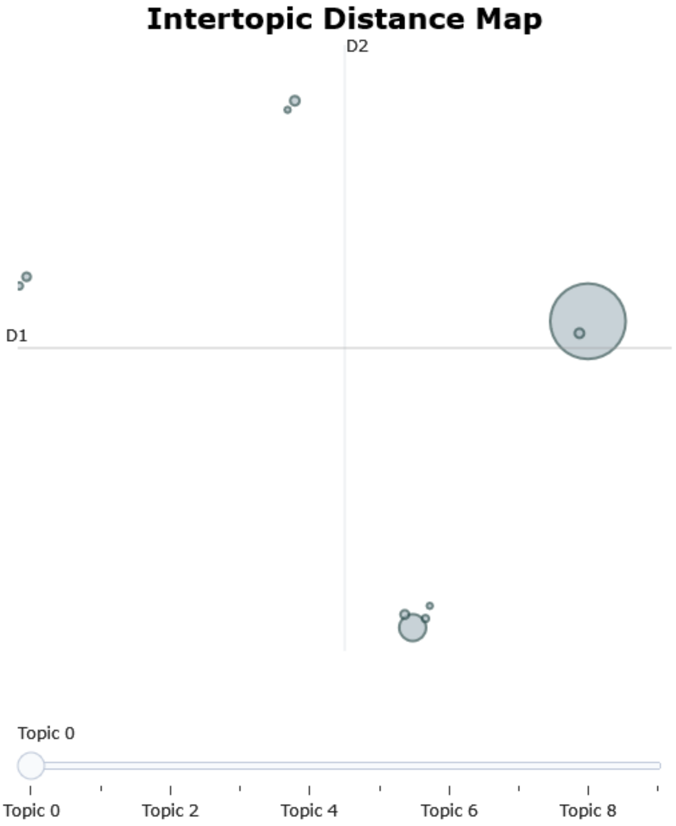


Figure 6: Intertopic Distances in StackOverflow (6k) ban Thread

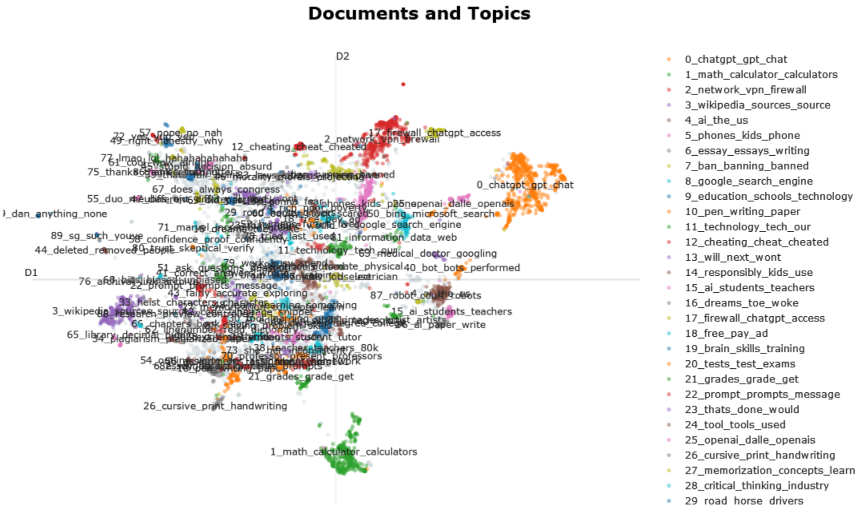


Figure 7: Documents and Topics in NYC ban Thread

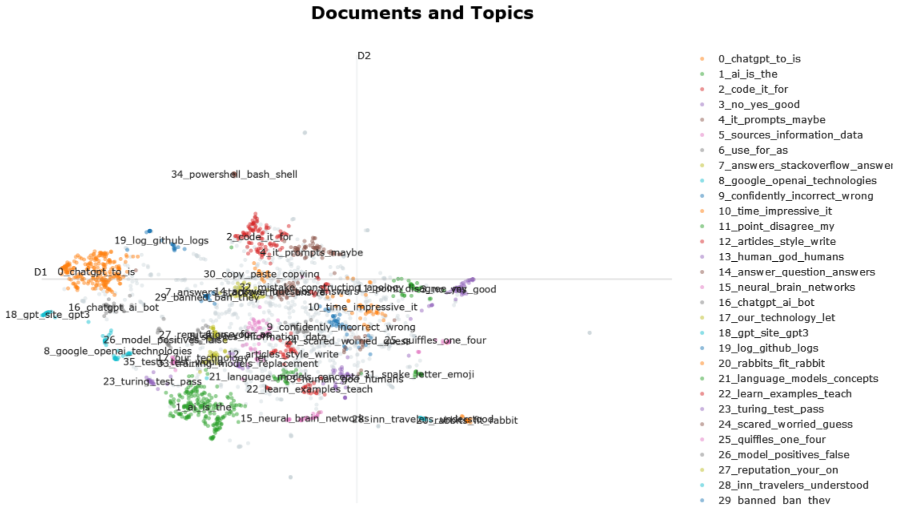


Figure 8: Documents and Topics in StackOverflow (6k) ban Thread

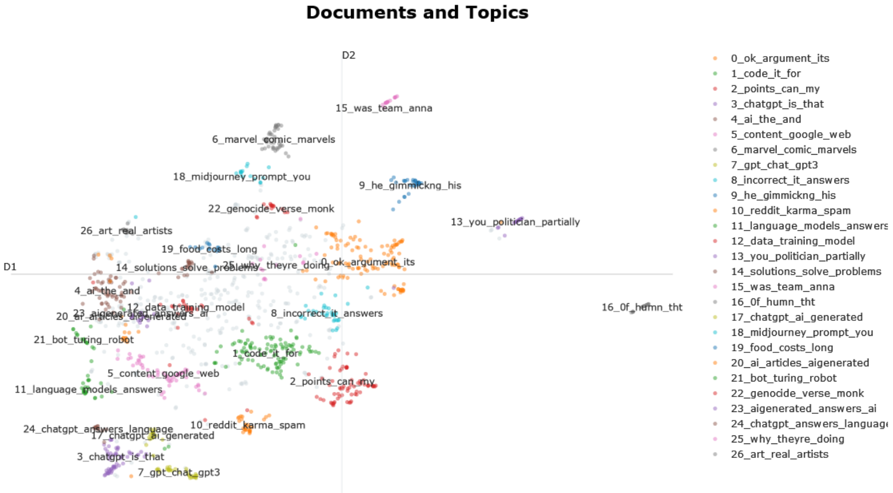
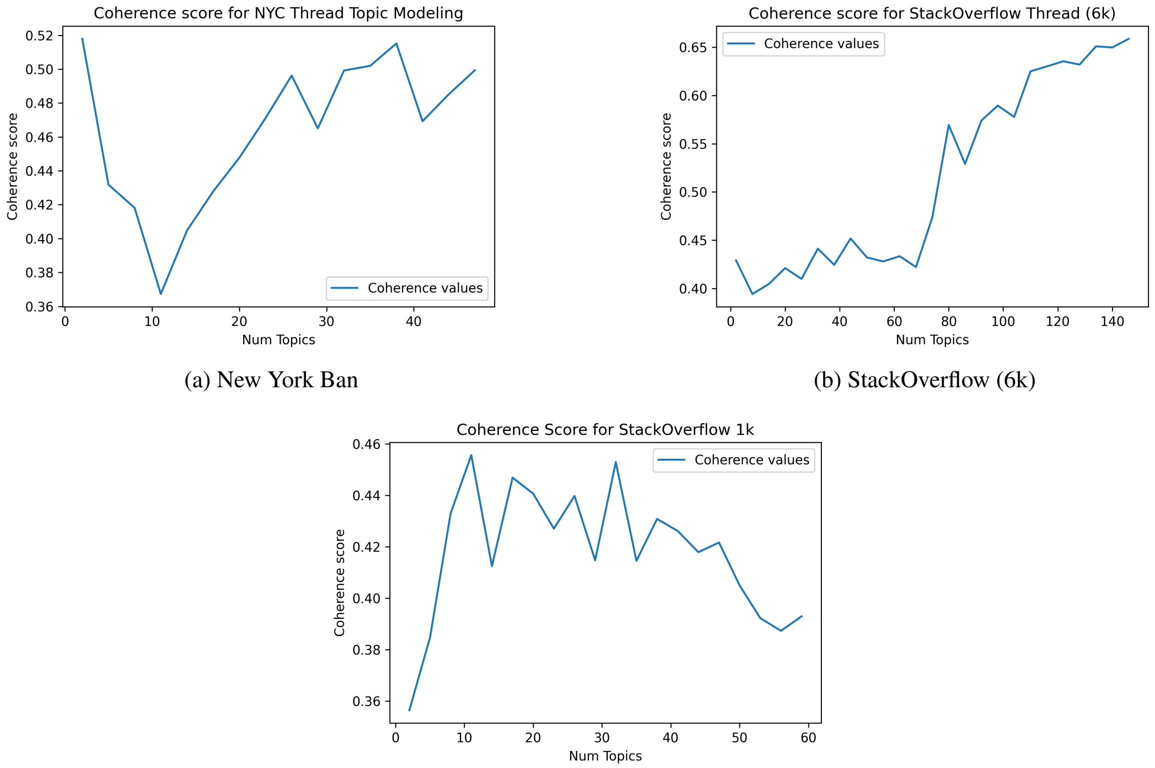


Figure 9: Documents and Topics in StackOverflow (1k) ban Thread

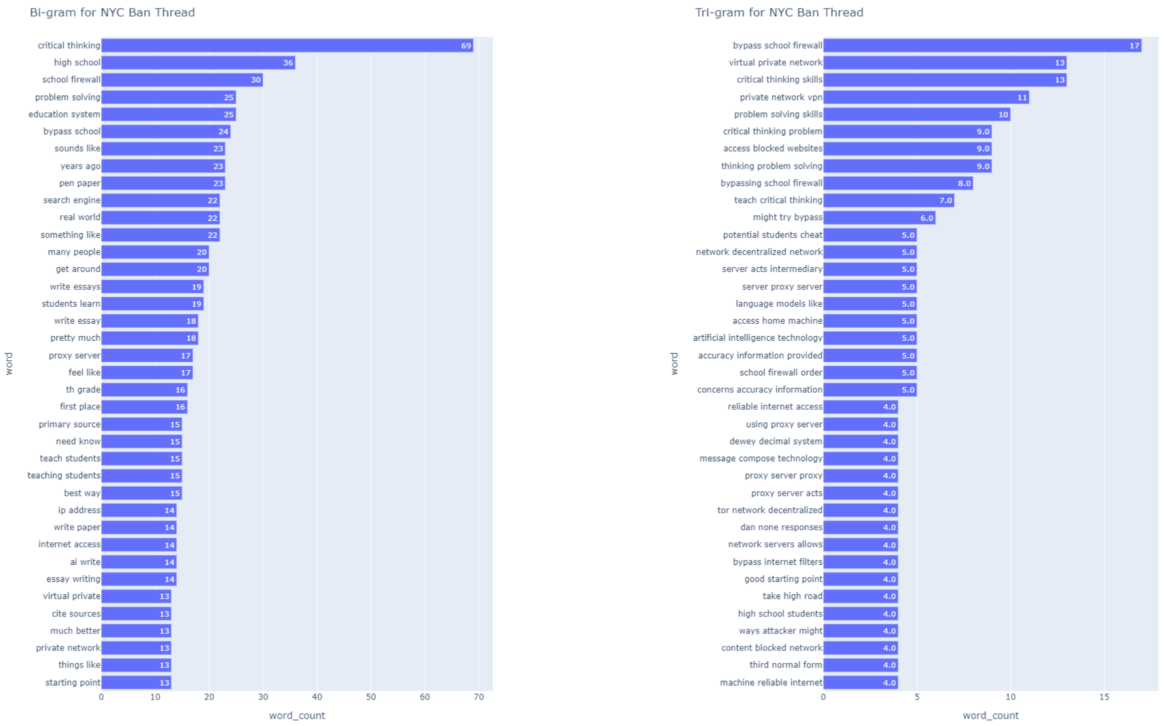


(a) New York Ban

(b) StackOverflow (6k)

(c) StackOverflow (1k)

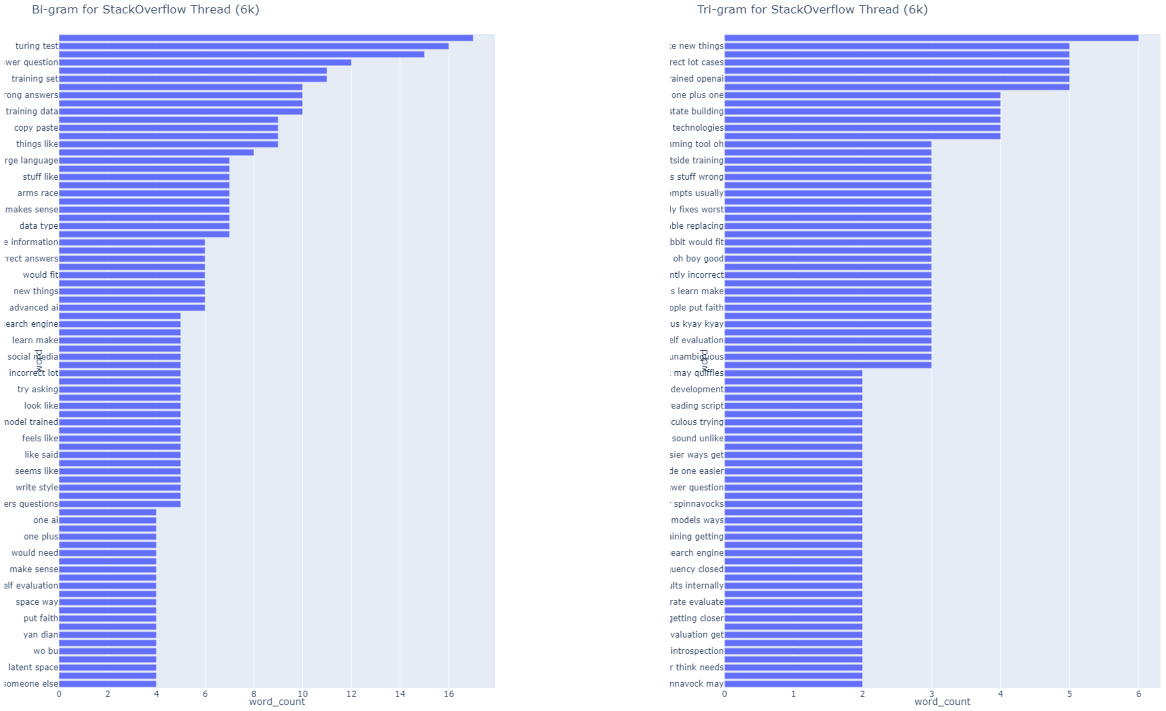
Figure 10: Coherence Scores



(a) Bigram

(b) Trigram

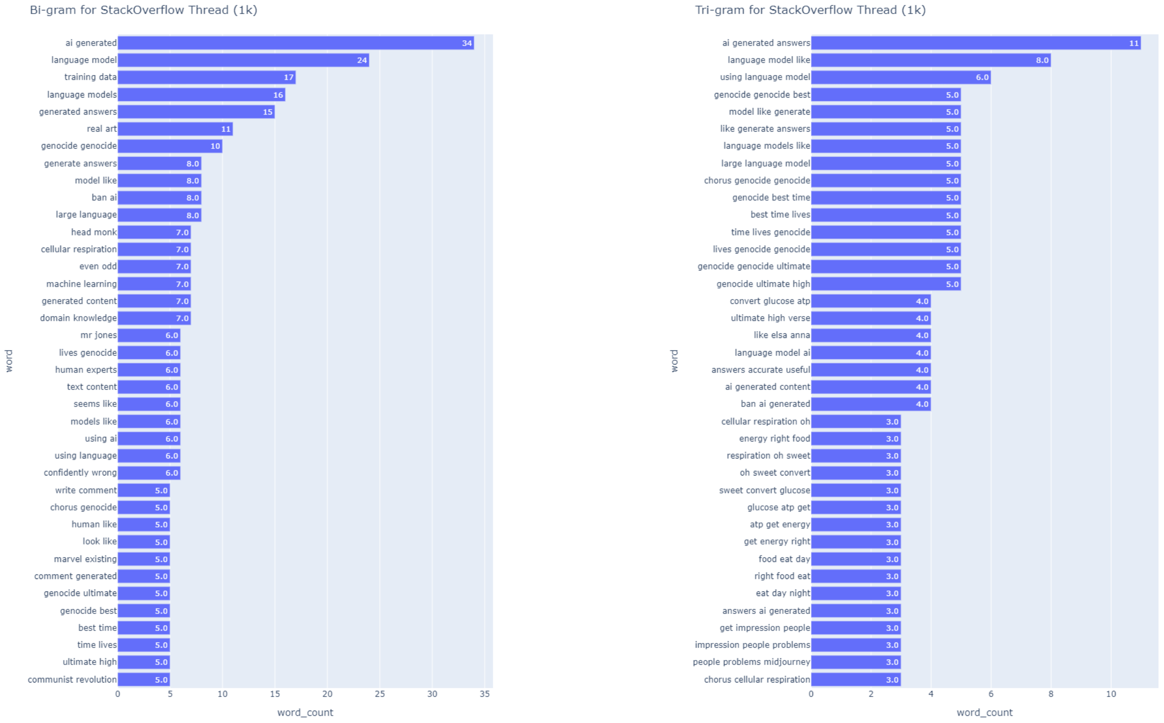
Figure 11: N-grams of Reaction to NYC Ban



(a) Bigram

(b) Trigram

Figure 12: N-grams of Reaction to StackOverflow Ban (6k)



(a) Bigram

(b) Trigram

Figure 13: N-grams of Reaction to StackOverflow Ban (1k)

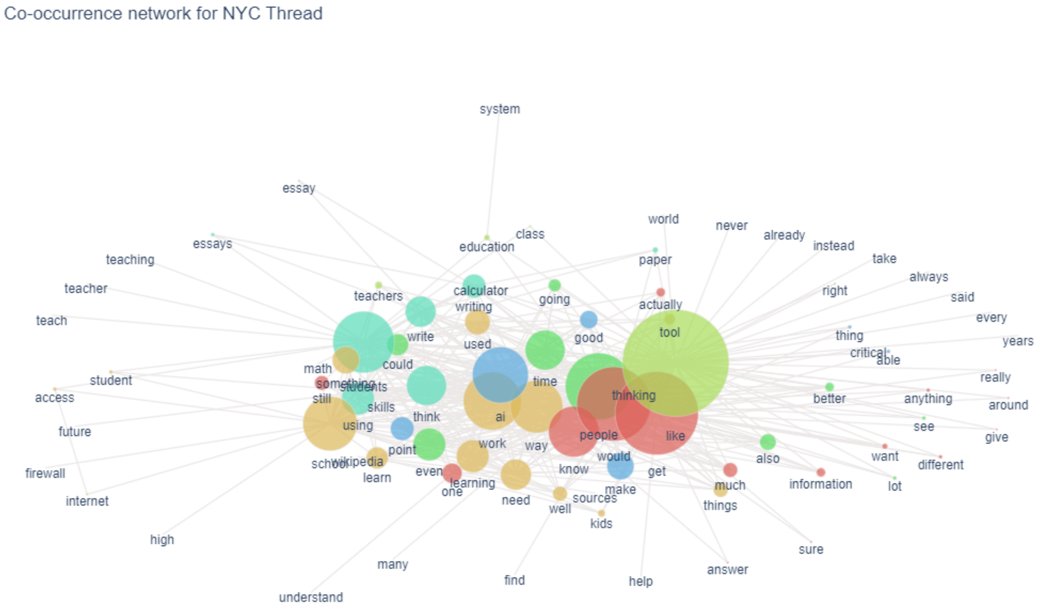


Figure 14: Co-occurrence network of keywords in NYC ban Thread

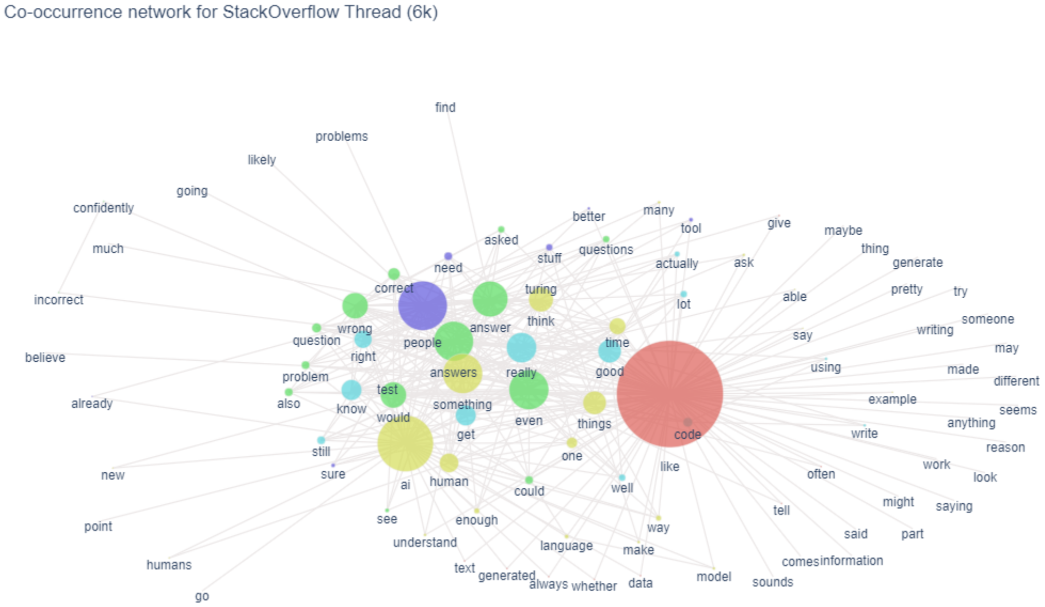


Figure 15: Co-occurrence network of keywords in StackOverflow ban Thread (6k)

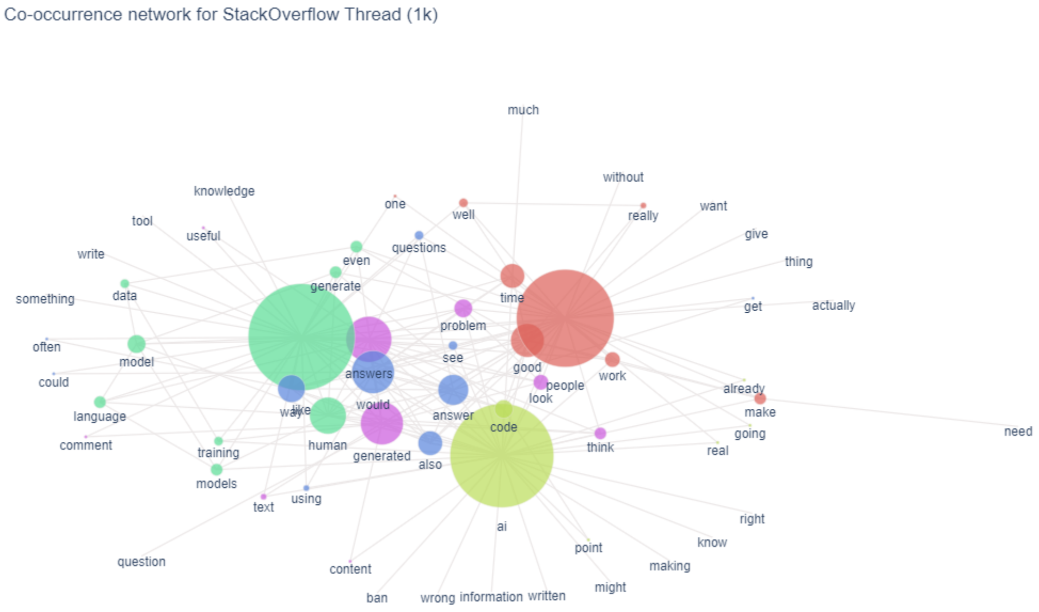


Figure 16: Co-occurrence network of keywords in StackOverflow ban Thread (1k)

Girlbosses, The Red Pill, and the Anomie and Fatale of Gender on Social Media: Analyzing Posts from r/SuicideWatch on Reddit

Elissa Nakajima Wickham
Waseda University*
elissa@akane.waseda.jp

Abstract

The proliferation of social media use in daily life has introduced a new practice in today's society: posting about suicidal ideation or intent online. Recent trends in social media reflect a movement towards different forms of male and female empowerment that impact gender norms, and thus, may impact social categorization. This pilot study explores posts from r/SuicideWatch that include discussions of gender and its implications for online conceptions of social identity. We use computational methods borrowed from natural language processing to analyze this impact from a novel perspective rarely seen in sociology.

1 Introduction

There is an estimated population of 4.8 billion social media users as of 2022 (Beltran et al., 2022). Several studies have identified the role of social media in the construction of group identity, from college sports (Kim and Kim, 2019) to political identity (Bennett, 2012). Further, social comparison through social media has been researched in relation to the impact on identity and self-esteem (Vogel et al., 2014). Expressing suicidal intent on social media occurs often enough to be explored as a public health issue (Luxton et al., 2012), and there are several different communities online dedicated to providing peer support and a place to share for those suffering from suicidal ideation.

This pilot study utilizes the findings of Emile Durkheim's *Suicide: A Study in Sociology*(Durkheim, 2005) to explore present-day motivations for suicide through the analysis of posts by users of Reddit concerning their own suicidality. The emergence of new models online for the "empowered" woman or man, paired with the social disruption during the COVID-19 pandemic, necessitates an inquiry into how changing gender norms may or may not be distressing for individuals. This study is based on a BERTopic model of text posts from r/SuicideWatch, a subreddit for peer-support for suicidal ideation. This paper will discuss the topics identified in the model that coincide with Durkheim's theory, especially as they relate to recent narratives on social media surrounding empowerment and gender.

2 Theoretical Basis

2.1 Durkheim's *Suicide: A Study in Sociology*

It is first necessary to briefly summarize Durkheim's findings in his 1897 book, *Suicide: A Study in Sociology* (Durkheim, 2005), before discussing its potential relevance to any topics in the posts on r/SuicideWatch. The key theoretical constructs from his work utilized in this study relate to his findings on social integration and moral regulation (Mueller et al., 2021). An imbalance in either creates four types of suicide, as visualized in the table on the following page.

The two factors that impact occurrences of suicide are regulation and integration, with the latter having a stronger impact according to Durkheim's findings (Mueller et al., 2021). Moral regulation has to do with whether the social rules of a particular group are made clear to an individual, and the degree to which the individual feels pressure from them (Mueller et al., 2021). Too little regulation, where an individual loses touch with the guiding force of morality in the group, usually in times of rapid social change, results in anomic suicide. On the other hand, too much regulation, where an individual may feel too much pressure or coercion as the result of the demands for a particular social group results in fatalistic suicide.

The norms of a group expressed as morality provide a certain protection for individuals, and the same is true for Durkheim when it comes to social integration. Durkheim argues that social bonds

*MA Candidate at the Graduate School of Asia-Pacific Studies

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 195–212
December 1-3, 2023 ©2023 Association for Computational Linguistics

Cause of Suicide	Type of Suicide
> Moral Regulation	1. Fatalistic Suicide
< Moral Regulation	2. Anomic Suicide
< Social Integration	3. Egoistic Suicide
> Social Integration	4. Altruistic Suicide

Table 1: Four Types of Suicide and Causes according to Durkheim

serve to protect individuals from the sufferings associated with life (Durkheim, 2005, 209-210). He understands social integration as the extent and depth of an individual's social relationships (Mueller et al., 2021). Thus, too little integration results in egoistic suicide, in which an individual feels a lack of belonging with a group, and loses the bond that would otherwise protect them from distress. Too much integration results in altruistic suicide, which sees an individual complete the act as a result of too much pressure through social bonds leading to the conclusion that their death would be for the best in terms of the health of their group.

2.2 Gender as a Social Category

In this study, our use of the term gender is based on the concept of social categories. Social categories comprise a group of people defined by a label either given to or used by the group of people, and the label itself must be utilized often enough that it impacts larger society's thinking or behavior around the group (Fearon, 1999). Social categories are defined by two features. The first are the rules of membership either explicit or implicit that include or exclude people from the group. The second are the sets of characteristics or certain types of behaviors that are alleged to be common to or expected of the category. These rules of membership and these sets of characteristics are debated and may change over time, but ultimately are consequential in the way they influence and condition ways of thinking and patterns of behavior (Fearon, 1999).

Our two social categories for gender are the binary "man" and "woman." We are using this conceptual definition of gender for the reason that the rules of membership and sets of characteristics that inform a social category mirror Durkheim's understanding of the purpose of moral regulation for social groups: both remain beyond the control of the individual and create expectations for behavior and ways of being. To comment on the levels of social integration online and how they may impact suicidal ideation for men and women requires a separate task of determining how the different categories integrate through the internet. For this reason, this paper will focus on the relationship between moral regulation online, or the rules of membership and alleged sets of characteristics, and suicidality for men and women. It becomes possible to understand the expectations online of for men and woman as the social morality that would result in either fatalistic or anomic suicide.

3 Background

3.1 r/SuicideWatch

The online platform Reddit, with 52 million active users (Beltran et al., 2022), allows those with accounts, or "redditors" to submit user-generated content, like links, text posts, and images; to relevant "subreddits" or "sub," which are user-created boards centered around the discussion of a particular subject[1]. The subreddit of interest for this paper is r/SuicideWatch, a subreddit started in December of 2008 that claims to provide an online space for "anyone struggling with suicidal thoughts."[2] To date, the subreddit has 424,000 members.

This subreddit has been selected for analysis for two reasons. The first is that the number of members in the sub makes it one of the largest resources for anonymous user-generated content around suicidal ideation; and the second is that the guidelines of interacting with posters in the subreddit are conducive to honest self-reporting of factors behind their suicidality. Two important guidelines from the r/SuicideWatch's "Talking Tips" page are the prohibitions around giving advice and offering encouragement. The moderators discourage offering help, as it may make the suicidal person feel even more powerless, and further discourage any members from accidentally invalidating a poster's feelings by posting uplifting remarks.

3.2 Social Media and Online Gender Empowerment Movements

With the impact of social media on self image being well-explored in the literature (Vogel et al., 2014), it is worth examining the ways in which gender roles have evolved online. The term "Girlboss" was created by NastyGal founder Sophia Amoruso in 2014, "to describe a way of presenting a professionally successful persona that highlights femininity"

[1] https://en.wikipedia.org/wiki/Reddit

[2] https://www.reddit.com/r/SuicideWatch/

(Atir, 2022). “#GirlBoss” on Instagram has 27 million posts to date[3], and TikToks tagged “#Girlboss” have an aggregate of 10.3 billion views[4]. While this hashtag only represents one expression of female empowerment on social media, the overall trend has been identified as one that reinforces the concept of “meritocracy” under capitalism (Robinson, 2023). The concept of an empowered "girlboss" represents “an ‘ideal’ of a woman who has it all” (Robinson, 2023).

In terms of social media trends around male gender roles, an increasingly popular concept called the “Manosphere” has steadily gained traction over time, which includes concepts like the redpill, alpha males, and involuntary celibates (incels) (Ging, 2019). The philosophy of the “Manosphere” has been described by some scholars as, “superficially [resolving] a contradiction between hegemonic masculinity’s prescriptive emotional walls and an inherent desire for connection by constructing women as exchangeable commodities” (Van Valkenburgh, 2021). Like “#Girlboss”, “#Alphamale,” a term directly associated with manosphere beliefs has millions if not billions of views on social media; with 1 million views on Instagram[5] and 2.4 billion on TikTok[6].

Previous research on the difference in suicidality between men and women has identified interpersonal events as triggers for women and achievement events as triggers for men (Waelde et al., 1994). The concepts of “#Girlboss”, centered around achievement and success in the professional world, and “#Alphamale”, centered around success and achievement in interpersonal relationships as it relates to other men and relative success in finding romantic or sexual partners (Van Valkenburgh, 2021), pose rules of membership and sets of characteristics opposite to established triggers for men and women. The concepts of a “Girlboss” and an “Alphamale” on social media, thus, provide potential conflict within the social category of "man" and woman" on the internet.

4 Research Questions and Hypotheses

This leads us to the presentation of the research questions for this study, which can be understood as follows:

- Can online empowerment movements targeting men and women affect the rules and expected characteristics for the group as a whole?
- Are gender norms online overly confining (fatalistic) or ambiguous and poorly defined (anomic)?

For men and women posting in r/SuicideWatch, being overly confined by the expectations, or moral regulation, of their social category would lead to fatalistic suicidality; while being confused by moral regulations would lead to anomic suicidality. These questions then lead to three hypotheses to test in this paper:

I. If online empowerment movements can impact the morality of a social category, then we can expect to see topics related to each one in posts made by men and women in r/SuicideWatch.

II. If gender expectations online create an anomic environment, the we can expect to see topics related to overwhelming confusion or feeling lost.

III. If gender expectations online create a fatalistic environment, the we can expect to see topics related to feeling trapped or controlled.

5 Methodology

The data for this study is from a publicly available dataset on Kaggle. Utilizing the Pushshift API to scrape posts from r/depression and r/SuicideWatch for suicide and depression detection, this dataset contains 484,969 posts from r/SuicideWatch between 2008 and 2021.[7] To inquire more into the potential influence of online empowerment movements on gender and suicidality, we focused on posts in the Kaggle dataset posted after January 1st, 2016 up until 2021, as Google Trends reports that searches for "girlboss" peaked in March of 2016, while searches for "alphamale" has remained steady.[8] Because the gender identity of the posters is not specified in this particular dataset, we then isolated the posts that contained age and gender identifiers, typically denoted on the website as

[3] https://www.instagram.com/explore/tags/girlboss/?hl=en

[4] https://www.tiktok.com/tag/girlboss?lang=en

[5] https://www.instagram.com/explore/tags/alphamale/?hl=en

[6] https://www.tiktok.com/tag/alphamale?lang=en

[7] https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch

[8] https://trends.google.com/trends/explore?date=all&q=girlboss,alphamale

some variation of "30F" or "18m." This was done using regular expressions (RegEx) to match patterns of only two digits followed by a capital or lowercase "F" or "M" within the body of the texts. This led to the creation of a new dataset of 1381 posts containing female identifiers, another dataset of 4190 posts containing male identifiers, and a dataset of 301 posts containing both.

The two datasets with either a female identifier or a male identifier were then used to create two BERTopic models. BERTopic is a technique for topic modeling, utilizing c-TF-IDF for clustering and embedding models to form topics and facilitate easier interpretation while preserving important terms in the description for each topic (Grootendorst, 2022). Sentence transformers were then used to compute sentence embeddings within the text in order to compute semantic relationships between sentences in the text and further refine the models (Reimers and Gurevych, 2019). We then utilized Maximal Marginal Relevance (MMR) as the representation model to calculate MMR between candidate terms and the document itself. This approach considers the similarity of key terms between others and in the context of the entire document, resulting in better diversity within the extracted topics and their terms (Grootendorst, 2022). The final BERTopic model for the posts containing female identifiers yielded 278 topics, while the BERTopic model for the posts with male identifiers yielded 646 topics. The appendix contains 200 of the most salient topics with their count, name, and representation for each subset of the data.

6 Results: Topics in r/SuicideWatch

6.1 Hypothesis I: *If online empowerment movements can impact the morality of a social category, then we can expect to see topics related to each one in posts made by men and women in r/SuicideWatch.*

College and school performance are both included in topics from each dataset, the representation of this topic was 949 out of 1381 posts for women (Topic 1), while it was 528 out of 4190 for men (Topic 4). Failure and success are also represented in women (Topic 41), and men (Topic 42). Money is also discussed for women (Topic 27, 154), and for men (Topic 2). The presence of these topics in posts from r/SuicideWatch with female identifiers suggests the influence of the "GirlBoss" movement, but cannot be conclusive. It is important to note interpersonal events do appear in the women dataset, with references to infidelity (Topic 140) and family and relationship troubles (Topic 64, 80, 97, 124). However, topics concerning school and work, while may not conclusively suggest the influence of the Girlboss empowerment movement, does suggest that achievement can be a notable trigger for suicide in women, contrary to previous research. Less ambiguous is the influence of "Manosphere" or "Alphamale" ideas present in the topics for posts from r/SuicideWatch containing male identifiers. There are topics concerning being unattractive (Topic 44), facing rejection from women (Topic 58), being a virgin (Topic 157), and physical fitness (Topic 159). As previously discussed, achievement-related topics were present in the dataset, but it is also important to note that interpersonal events did appear in topics discussing conflict with parents (Topic 28) and breakups (Topic 169), but with less frequency.

6.2 Hypothesis II: *If gender expectations online create an anomic environment, the we can expect to see topics related to overwhelming confusion or feeling lost.*

For the women's dataset, topics related to confusion or feeling lost were more prevalent than in the men's dataset. This is discussed in topics covering lack of belonging (Topic 48), worthlessness(Topic 45), guilt and shame (Topic 63), being a burden (Topic 74), having nothing left (Topic 88), not knowing what the point of life is (Topic 104), and feeling numb (Topic 111). These topics were not absent from the men's dataset, represented in topics on guilt and shame (Topic 67), worthlessness (Topic 72), not knowing what's real (Topic 107), feeling empty (164), and being being a burden (110).

6.3 Hypothesis III: *If gender expectations online create a fatalistic environment, the we can expect to see topics related to feeling trapped or controlled.*

For the men's dataset, topics related to feeling trapped or controlled were more prevalent than in the women's dataset. These topics comprised representations of not wanting to live anymore (Topic 88), not wanting to be here anymore (Topic 92), reactions of anger and upset (Topic 129, 139), stress (Topic 141), control (Topic 167), not being able to stand life (Topic 174), being crushed (Topic 178), and exhaustion (Topic 139, 186). However, similar topics were also present in the women's dataset, although less frequently. These topics covered representations of not being able to take it anymore (Topic 132), life being too much (Topic 164), wanting to escape (Topic 166), not being able to handle life (Topic 182), and being tired (Topic 53).

7 Analysis

According to Fearon, social categories can be further subdivided into role and type categories, where performance of the role and and the principles associated with the role, or type, help individuals orient themselves within society (Fearon, 1999). While we have offered his definition of social identity through social category, he distinguishes social identity from personal identity, or "distinguishing features of a person that form the basis of his or her dignity or self-respect" (Fearon, 1999). It is through the interaction of social and personality identity that it becomes possible to understand moral regulation online and its impact on suicidality.

7.1 Anomie: GirlBoss as a Social Identity

The results of this study suggest that women may experience more anomie in their ideation, or disconnection from the moral regulation of their social category, which warrants a discussion of why this may be. The female-identifiers within the posts on r/SuicideWatch suggest a younger population of posters (Topic 4, 38). Girlboss joins other hashtags like AddWomen and STEMinism in encouraging women's participation and acceptance in the professional and academic realm. Because these empowerment movements involve the expansion of the social category of women to include more roles, like scientist or CEO, it can make it difficult for individuals within the social category to orient themselves within differing role expectations, especially for younger women. Gender discrimination in male-dominated spheres may also complicate type expectations for women performing different roles, as "masculine" behavior is seen as being more conducive to success in the field (Van Veelen et al., 2019). The ambigutity and presence of different role expectations may make it difficult for the individual to understand the alleged rules and characteristics they need for membership in their social category. This is not to disparage feminist movements online, but rather to point to how the necessity of tying the social category of woman to different roles in order to alleviate gender inequality, may have the unintended consequence of causing distress to individuals. As women are encouraged to take on roles previously unavailable to them, it is important to be attentive to how social pressures may result in an anomie where they feel lost among differing expectations.

7.2 Fatale: Alphamale as a Personal Identity

The suicidal ideation expressed in r/SuicideWatch for men leans fatalistic, and this is where personal identities become helpful in explaining this result. While Girlboss empowerment represents one social category (CEO) being merged into another (woman), the Alphamale links a personal identity to the social category of man. The distinguishing features of the Alphamale that comprise physical fitness, wealth, and the ability to have sexual success becomes the basis for the individual's self-esteem. For those who have absorbed the philosophy of the manosphere, sexual success not only becomes the basis of their dignity, but also the means through which they orient themselves within society. Moreover, the moral regulation resulting from this convergence is extremely confining. As noted in Van Valkenburgh (2021), those falling into the beliefs of the "manosphere" assert that, "human nature and behavior are essentially unchanging and rooted in biological determinants" (Van Valkenburgh, 2021). In other words, if a man is experiencing a lack of success in finding sexual partners, they must face the "truth" that their situation is both unchanging and biologically determined. Fearon argues that individuals with fewer social identities tend to hold personal identities more firmly (Fearon, 1999). Because the particular ideology of the manosphere is particularly dangerous, resulting in mass murder-suicides known as "misogynistic terrorism"[9], it may be valuable to focus on strengthening and furthering men's social identities. That way, personal identities may become less attractive for the individual and the rules of membership and sets of characteristics may feel less confining, as the personal identity no longer serves to orient the individual within society.

Limitations

This paper is not intended to replace medical and psychological research on suicide, nor should the findings of this paper ever be prioritized over clini-

[9] https://en.wikipedia.org/wiki/Misogynist_terrorism

cal expertise. Our intention is to identify potential sources of distress for individuals participating in the online space. While user-generated content on suicidal ideation can provide insight into the possible social motivations for suicide, they do not offer tangible data on whether or not the act is completed. The anonymity provided by Reddit posts, as users are not required to provide their full name or image in creating an account, complicates any possible determination of whether the ideation expressed in their posts translates into action in real life. Therefore, this study should not be used as a basis for treatment or prevention of suicide.

References

Stav Atir. 2022. Girlboss? highlighting versus downplaying gender through language. *Trends in Cognitive Sciences*.

Cynthia Beltran, Maria Jacal, Kendra Saucedo, Jose Mendez, and Alvaro Zuaznabar. 2022. Stats vs. facts.

W Lance Bennett. 2012. The personalization of politics: Political identity, social media, and changing patterns of participation. *The annals of the American academy of political and social science*, 644(1):20–39.

Emile Durkheim. 2005. *Suicide: A study in sociology*. Routledge.

James D. Fearon. 1999. What is identity (as we now use the word)?

Debbie Ging. 2019. Alphas, betas, and incels: Theorizing the masculinities of the manosphere. *Men and masculinities*, 22(4):638–657.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Bumsoo Kim and Yonghwan Kim. 2019. Growing as social beings: How social media use for college sports is associated with college students' group identity and collective self-esteem. *Computers in Human Behavior*, 97:241–249.

David D Luxton, Jennifer D June, and Jonathan M Fairall. 2012. Social media and suicide: a public health perspective. *American journal of public health*, 102(S2):S195–S200.

Anna S Mueller, Seth Abrutyn, Bernice Pescosolido, and Sarah Diefendorf. 2021. The social roots of suicide: Theorizing how the external social world matters to suicide and suicide prevention. *Frontiers in psychology*, 12:763.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Leone Robinson. 2023. Is the 'girl boss' really just an anti-feminist commodification? *Critical Reflections: A Student Journal on Contemporary Sociological Issues*.

Shawn P Van Valkenburgh. 2021. Digesting the red pill: Masculinity and neoliberalism in the manosphere. *Men and masculinities*, 24(1):84–103.

Ruth Van Veelen, Belle Derks, and Maaike Dorine Endedijk. 2019. Double trouble: How being outnumbered and negatively stereotyped threatens career outcomes of women in stem. *Frontiers in psychology*, 10:150.

Erin A Vogel, Jason P Rose, Lindsay R Roberts, and Katheryn Eckles. 2014. Social comparison, social media, and self-esteem. *Psychology of popular media culture*, 3(4):206.

Lynn C Waelde, Louise Silvern, and William F Hodges. 1994. Stressful life events: Moderators of the relationships of gender and gender roles to self-reported depression and suicidality among college students. *Sex Roles*, 30:1–22.

A Appendix

Table 2: Topics from dataset of posts containing female identifiers

Topic	Count	Name
-1	9314	-1_he_my_him_and
0	1630	0_she_her_shes_we
1	949	1_job_school_college_degree
2	476	2_suicidal_suicide_thoughts_attempted
3	295	3_friends_friend_have_friendships
4	285	4_19f_18f_15f_21f
5	224	5_kill_myself_killing_die
6	223	6_sleep_wake_bed_waking
7	217	7_sorry_post_read_reading
8	190	8_happy_happiness_joy_miserable
9	182	9_pills_meds_overdose_take
10	176	10_pain_hurts_hurt_painful
11	175	11_ugly_look_mirror_face
12	168	12_mom_mother_mum_she
13	167	13_alone_lonely_isolated_feel
14	166	14_die_want_dead_wish
15	155	15_them_they_their_tell
16	152	16_abused_raped_sexually_abuse
17	146	17_depression_depressed_years_been
18	144	18_family_theyre_them_my
19	142	19_hate_myself_hates_hating
20	139	20_anxiety_panic_attacks_attack
21	135	21_felt_feel_feels_feeling
22	131	22_depression_anxiety_19f_23f
23	128	23_cry_crying_cried_sobbing
24	126	24_parents_siblings_my_tell
25	125	25_bridge_jumping_edge_jump
26	119	26_therapist_therapy_therapists_tried
27	117	27_bills_savings_debt_money
28	106	28_we_together_weve_each
29	104	29_eating_eat_weight_food
30	103	30_gun_pistol_guns_rounds
31	102	31_cat_cats_dog_pets
32	101	32_suicidal_suicide_thoughts_past

Continued on next page

Table 2: Topics from dataset of posts containing female identifiers (Continued)

Topic	Count	Name
33	99	33_end_ending_it_want
34	94	34_dad_father_abusive_he
35	89	35_care_cares_cared_about
36	89	36_thoughts_dark_intrusive_these
37	86	37_love_loves_never_someone
38	85	38_17f_20f_16f_25f
39	84	39_room_quarantine_apartment_clean
40	81	40_care_me_anyone_life
41	80	41_failure_fail_failed_failing
42	79	42_him_he_film_hes
43	78	43_self_harm_harming_selfharm
44	76	44_hobbies_interest_pointless_games
45	75	45_worthless_useless_worth_feel
46	72	46_live_living_reason_inspiring
47	72	47_hospital_patients_inpatient_hospitalization
48	70	48_here_belong_be_still
49	68	49_bullied_school_bullying_grade
50	61	50_police_cops_report_called
51	61	51_help_ask_professional_need
52	57	52_life_sick_hate_isnt
53	55	53_tired_im_nnim_so
54	55	54_better_get_gets_wellnnnbut
55	55	55_student_college_training_school
56	55	56_what_do_know_dont
57	54	57_nnfast_forward_today_months
58	53	58_fight_fighting_flight_surrender
59	53	59_drink_sober_drank_drinking
60	51	60_stop_want_away_stopped
61	50	61_died_dead_funeral_death
62	49	62_cutting_cut_again_clean
63	49	63_guilt_guilty_shame_forgive
64	49	64_he_him_hes_upset
65	48	65_advice_any_appreciate_anyone
66	48	66_drugs_weed_marijuana_smoke
67	48	67_brother_close_cousin_little
68	48	68_disorder_personality_borderline_diagnosed

Continued on next page

Table 2: Topics from dataset of posts containing female identifiers (Continued)

Topic	Count	Name
69	47	69_passed_died_grandmother_cancer
70	47	70_covid_pandemic_economy_covid19
71	46	71_ambulance_hospital_cabinet_nurse
72	44	72_scared_terrified_worried_afraid
73	44	73_hi_hello_hey_everyone
74	44	74_burden_everyone_perform_ruining
75	43	75_anymore_cant_do_apparentlyni
76	43	76_god_pray_religious_prayed
77	43	77_nncheers_nnthanks_nna_accurate
78	43	78_plan_plans_planner_backup
79	42	79_mental_health_illness_mentally
80	42	80_children_kids_married_points
81	40	81_bipolar_bpd_diagnosed_colleague
82	40	82_hope_excited_hopeful_no
83	40	83_lesbian_men_incel_bisexual
84	40	84_he_told_him_didnt
85	39	85_hurt_hurting_anyone_grieve
86	39	86_want_dont_sugarcoat_request
87	39	87_breathe_chest_oxygen_breath
88	38	88_nothing_anywhere_left_hold
89	38	89_selfish_jealous_selfishnntoday_respectfully
90	37	90_broken_broke_apart_spirit
91	37	91_future_bleak_see_fishies
92	37	92_fat_weigh_pounds_overweight
93	37	93_strong_weak_stronger_strength
94	36	94_dreams_dream_dreaming_crushed
95	36	95_one_no_nnat_anyone
96	36	96_19f_lovedcared_forwanted_relationship
97	35	97_texted_messaged_texting_him
98	34	98_age_aged_youngest_child
99	34	99_sick_nauseous_stomach_throw
100	34	100_problems_conflicts_fault_excuse
101	33	101_vent_venting_needed_messy
102	33	102_parents_accomplish_therapist_therapy
103	33	103_pregnant_abortion_lost_baby
104	32	104_point_whats_continuing_see

Continued on next page

Table 2: Topics from dataset of posts containing female identifiers (Continued)

Topic	Count	Name
105	32	105_know_knows_clue_dont
106	32	106_account_throwaway_deleted_main
107	32	107_would_brains_woods_drink
108	31	108_tried_attempt_times_noose
109	31	109_understand_understandable_understands_problematic
110	30	110_rgangstalking_udrunkenposting_shielding_mod
111	30	111_numb_numbness_hollow_dumbness
112	30	112_landlord_eviction_landlords_rental
113	30	113_worked_works_nothing_work
114	30	114_car_driving_mph_freeway
115	30	115_person_bad_promise_assuming
116	30	116_ocd_experiences_rational_explanations
117	29	117_better_get_harder_struggling
118	29	118_drowning_drown_afloat_drowned
119	29	119_done_sooooo_piling_board
120	29	120_worse_gotten_gets_getting
121	28	121_peace_heavennnampx200bnntldr_patiently_gates
122	28	122_talk_one_no_anyone
123	28	123_existence_living_existing_nnnone
124	28	124_he_abusive_sociopath_inexplicable
125	28	125_do_nnwhat_what_should
126	28	126_night_day_sometimes_ntonight
127	28	127_decades_its_fed_been
128	27	128_over_nnrant_ready_matrix
129	27	129_yeah_yay_nope_betcha
130	27	130_am_notni_arent_stating
131	26	131_cant_win_simply_theoretically
132	26	132_take_anymore_cant_longernnneed
133	26	133_sister_loves_sis_act
134	25	134_options_option_choice_choices
135	25	135_disgusting_ashamed_embarrassed_disgusted
136	25	136_media_followers_social_facebook
137	25	137_normal_pushing_deepdown_hints
138	25	138_die_want_genuinely_ready
139	25	139_exhausted_shaking_exhausting_resting
140	24	140_cheated_unfaithful_cheating_cheat

Continued on next page

Table 2: Topics from dataset of posts containing female identifiers (Continued)

Topic	Count	Name
141	24	141_hotline_prevention_lifeline_crisis
142	24	142_lie_lied_lying_questionnna
143	24	143_88f_btw_17f_intro
144	24	144_regret_service_ton_brother
145	24	145_thought_thinking_think_bingo
146	24	146_memory_remember_memories_vividly
147	23	147_death_fear_scared_scare
148	23	148_story_stories_share_strangers
149	23	149_changed_change_changes_reverted
150	23	150_deserve_deeds_deserved_bettered
151	23	151_others_likes_people_worrying
152	23	152_coward_courage_loyal_cowardnni
153	23	153_hair_shower_clumps_falling
154	23	154_function_progress_unstable_barely
155	23	155_born_wish_chose_never
156	23	156_talk_need_someone_please
157	23	157_disappeared_gone_vanish_notice
158	22	158_kissed_sex_sexual_penetration
159	22	159_empty_void_inside_feeling
160	22	160_surgery_hysterectomy_endometriosis_pelvic
161	22	161_condoms_condom_sex_protection
162	22	162_day_harder_struggle_everyday
163	22	163_smart_iq_intelligence_stupid
164	22	164_too_much_enough_its
165	22	165_blame_worst_games_video
166	21	166_escape_prison_ensured_fleeing
167	21	167_listening_heard_listen_nnthank
168	21	168_who_know_anna_unimportantnnanyways
169	21	169_hang_noose_hanging_tree
170	21	170_hate_english_sucks_pronounce
171	21	171_drawings_crossed_disgusted_shotaloli
172	20	172_tested_hiv_std_herpes
173	20	173_way_out_theres_see
174	20	174_social_skills_socially_needless
175	20	175_give_up_appreciatennthank_timennany
176	20	176_need_help_desperate_please

Continued on next page

Table 2: Topics from dataset of posts containing female identifiers (Continued)

Topic	Count	Name
177	20	177_deeper_lows_lower_low
178	20	178_ward_psych_councillor_psyc
179	20	179_worth_living_worthy_life
180	19	180_space_waste_clutter_wasting
181	19	181_art_passion_drawing_writer
182	19	182_handle_deal_cannot_anymore
183	19	183_mom_kill_mother_herself
184	19	184_fuck_shit_this_kind
185	19	185_universe_end_ending_yanking
186	19	186_emotions_positive_negative_mechanic
187	19	187_he_suicidal_him_doesnt
188	19	188_anymore_what_do_know
189	18	189_doctor_doctors_drs_diagnosis
190	18	190_motivation_motivated_recent_bottom
191	18	191_16_pocd_kill_pedophile
192	18	192_break_catch_crack_breaking
193	18	193_hated_tbst_bad_twenty
194	18	194_friend_friendship_group_immigrant
195	18	195_sad_angry_mad_driving
196	18	196_him_he_hes_3050
197	18	197_hell_sin_hellnnthen_hellcoaster
198	18	198_weather_90f_socks_degrees

Table 3: Topics from dataset of posts containing male identifiers

Topic	Count	Name
-1	27401	-1_her_she_and_my
0	2025	0_he_him_hes_his
1	1263	1_suicide_suicidal_thoughts_commit
2	761	2_job_debt_jobs_money
3	576	3_friends_group_friend_hang
4	528	4_grades_classes_college_exams
5	498	5_pain_hurt_hurts_hurting
6	477	6_kill_killing_myself_yourself
7	441	7_sleep_wake_bed_asleep
8	410	8_stomach_sick_shaking_nauseous

Continued on next page

Table 3: Topics from dataset of posts containing male identifiers (Continued)

Topic	Count	Name
9	405	9_happy_happiness_joy_excited
10	368	10_depression_depressed_severe_depressive
11	334	11_cry_crying_cried_tears
12	327	12_dose_dosage_50mg_500mg
13	317	13_therapist_therapy_therapists_seeing
14	293	14_care_cares_cared_about
15	293	15_mom_mother_her_moms
16	279	16_end_ending_tonight_want
17	261	17_better_worse_gets_things
18	250	18_gun_shooting_pistol_shotgun
19	236	19_car_road_tree_driving
20	223	20_thoughts_mind_head_these
21	200	21_family_care_cares_supportive
22	192	22_drinking_drink_drunk_drank
23	184	23_pills_pill_67_downed
24	184	24_dad_cant_just_dont
25	177	25_meds_medication_medications_medicine
26	175	26_drugs_heroin_drug_cocaine
27	171	27_weed_smoking_smoke_smoked
28	167	28_parents_blame_yell_love
29	164	29_panic_anxiety_attacks_attack
30	162	30_depression_depressed_dealing_years
31	162	31_feeling_feel_feelings_way
32	159	32_alone_isolated_feel_alonennim
33	158	33_2019_january_ago_march
34	157	34_old_22_age_27
35	153	35_psych_ward_nhs_psychiatric
36	145	36_she_told_her_asked
37	142	37_she_her_herself_shes
38	140	38_hate_hatred_myself_hating
39	139	39_fight_fighting_fights_fought
40	134	40_method_attempts_attempt_tried
41	134	41_we_together_each_both
42	130	42_failure_fail_failed_failing
43	130	43_father_dad_stepmom_mother
44	125	44_ugly_look_attractive_literal

Continued on next page

Table 3: Topics from dataset of posts containing male identifiers (Continued)

Topic	Count	Name
45	122	45_bridge_jump_jumping_building
46	119	46_bullied_bullying_bully_bullies
47	119	47_19m_16m_20m_17m
48	118	48_religious_god_religion_faith
49	118	49_awkward_social_socially_introverted
50	118	50_post_posted_posting_reddit
51	117	51_overdose_overdosed_overdoses_poisoning
52	115	52_cutting_cut_cuts_deeper
53	112	53_xanax_alcohol_bars_vodka
54	111	54_quetiapine_gt_clonazepam_trazodone
55	111	55_voice_voices_hear_scream
56	111	56_abused_raped_sexually_abuse
57	110	57_harm_self_harmed_harming
58	105	58_girls_rejected_rejection_girl
59	105	59_rent_apartment_pay_move
60	102	60_eat_food_eating_meals
61	101	61_27m_15m_19m_14m
62	97	62_antidepressants_anti_antidepressant_ssri
63	97	63_do_doing_itnnthings_wimp
64	91	64_paracetamol_500mg_paracetamols_tablets
65	89	65_dog_cat_cats_dogs
66	88	66_her_shes_hers_love
67	86	67_guilt_guilty_ashamed_shame
68	86	68_advice_appreciate_appreciated_any
69	86	69_thats_true_movies_alright
70	85	70_suicidal_suicide_thoughts_17m
71	83	71_prozac_20mg_40mg_took
72	82	72_worthless_worthlessness_worth_value
73	82	73_hi_hello_hey_reddit
74	79	74_help_assistance_seek_ask
75	79	75_plan_plans_planned_planning
76	78	76_peace_peaceful_sence_finally
77	77	77_mental_health_illness_stigma
78	76	78_future_see_ahead_luck
79	76	79_zoloft_100mg_50mg_200mg
80	76	80_account_throwaway_obvious_reasons

Continued on next page

Table 3: Topics from dataset of posts containing male identifiers (Continued)

Topic	Count	Name
81	76	81_birthday_birthdays_friday_21st
82	75	82_motivation_motivated_lazy_motivate
83	75	83_lie_lying_lied_lies
84	75	84_story_stories_share_read
85	74	85_nnfuck_nnok_nxo_nnyeah
86	73	86_read_reading_thank_reads
87	72	87_weigh_pounds_lbs_53
88	72	88_live_anymore_want_living
89	72	89_gay_bisexual_trans_sexuality
90	72	90_should_do_what_nnwhat
91	71	91_fear_scared_afraid_death
92	71	92_here_anymore_want_be
93	71	93_talk_anyone_speak_nobody
94	70	94_vodka_rum_bottle_alcohol
95	70	95_she_hand_called_stepping
96	69	96_memory_forget_memories_forgotten
97	69	97_hang_hanging_myself_tried
98	68	98_hobbies_interests_pursue_passion
99	67	99_start_title_says_protocall
100	67	100_hospitalized_hospital_jews_processed
101	66	101_scared_fear_afraid_terrified
102	66	102_ambulance_police_officers_paramedics
103	66	103_goodbye_bye_farewell_say
104	65	104_happen_happens_wieght_happened
105	65	105_surgery_hip_mri_spine
106	65	106_fault_blame_themselves_blaming
107	65	107_real_fake_feels_genuine
108	64	108_emotions_emotion_emotional_described
109	63	109_games_video_game_playing
110	63	110_burden_burdened_valuable_everyone
111	62	111_brother_older_younger_brothers
112	62	112_vent_venting_needed_guess
113	62	113_weak_strong_strength_weaker
114	61	114_suffering_suffer_continue_end
115	60	115_knife_blade_razor_sharp
116	60	116_seroquel_xr_25mg_50mg

Continued on next page

Table 3: Topics from dataset of posts containing male identifiers (Continued)

Topic	Count	Name
117	59	117_hospital_icu_rushed_released
118	59	118_deserve_entitled_rewarded_deserved
119	58	119_pandemic_covid_covid19_coronavirus
120	58	120_smart_intelligent_smarter_iq
121	58	121_trust_trusted_issues_trusting
122	57	122_music_guitar_songs_instruments
123	56	123_selfish_selfishnmy_breeding_selfishness
124	56	124_ibuprofen_800mg_200mg_aspirin
125	56	125_siblings_sister_sisters_youngest
126	56	126_focus_distract_concentrate_distracted
127	55	127_born_wish_yayyy_ethan
128	55	128_helped_helps_help_helping
129	54	129_upset_furious_shocked_shock
130	54	130_meant_joke_toonice_goofball
131	54	131_sad_sadness_breaches_somber
132	54	132_charged_court_probation_lawyer
133	54	133_taking_them_take_stopped
134	53	134_bipolar_disorder_manic_diagnosed
135	53	135_sertraline_150mg_50mg_100mg
136	52	136_bpd_adhd_anxiety_ocd
137	52	137_problems_problem_solve_temporary
138	52	138_did_woop_task_didnt
139	51	139_angry_anger_untreated_ridiculous
140	51	140_die_want_attentionnnive_consciencenn
141	50	141_stress_stressed_stressful_cohesive
142	50	142_fuck_wow_fucking_damn
143	49	143_text_texted_responded_glanced
144	49	144_understand_understands_they_societynnfor
145	49	145_confidence_selfesteem_mirror_esteem
146	49	146_anxiety_depression_severe_diagnosed
147	49	147_regret_regretted_regrets_decision
148	48	148_his_mom_he_him
149	48	149_brenda_claire_ruth_nate
150	48	150_ending_end_life_effective
151	48	151_pregnant_shes_she_pregnancy
152	48	152_fucked_up_fuck_someonensomething

Continued on next page

Table 3: Topics from dataset of posts containing male identifiers (Continued)

Topic	Count	Name
153	48	153_door_hallway_lock_locked
154	47	154_male_female_indian_17
155	47	155_took_take_more_20
156	47	156_rope_extension_hang_cord
157	47	157_virgin_virginity_hookers_indulge
158	46	158_vision_blind_eyes_blurry
159	46	159_gym_workout_routine_workouts
160	45	160_21m_creep_animated_acted
161	45	161_enjoy_fun_enjoyment_enjoyed
162	45	162_ptsd_coping_deployment_mechanisms
163	45	163_note_notes_wrote_written
164	44	164_empty_shell_hollow_space
165	44	165_sucks_bad_awful_garbage
166	44	166_scars_scar_wrist_thickest
167	44	167_control_controlling_controlled_selfcontrol
168	44	168_hope_hopeful_fading_hopes
169	44	169_cheated_girlfriend_digger_dumped
170	44	170_easier_easy_hard_idiotic
171	44	171_support_discord_plz_system
172	43	172_point_whats_gore_tremendous
173	43	173_parents_money_usd_dad
174	43	174_stand_take_anymore_cant
175	43	175_hopeless_hopelessnessnn_epitome_flow
176	43	176_need_help_please_posti
177	42	177_adhd_diagnosed_immediantly_mis
178	42	178_crushed_ordeal_made_safe
179	42	179_oh_yay_yes_yeah
180	42	180_stop_justjustjust_repelled_addy
181	42	181_talk_someone_need_please
182	41	182_they_them_believed_jesus
183	41	183_want_emo_dont_act
184	41	184_rambling_sorry_ramblings_ramble
185	41	185_social_anxiety_interacting_pressurized
186	41	186_exhausted_tired_blah_tiredness
187	41	187_pull_reporting_out_stuck
188	41	188_grades_studying_she_college

Continued on next page

Table 3: Topics from dataset of posts containing male identifiers (Continued)

Topic	Count	Name
189	40	189_miserable_adapts_chill_howling
190	40	190_teacher_teachers_class_joinery
191	40	191_do_scratch_really_dont
192	40	192_lexapro_escitalopram_10mg_15mg
193	40	193_hotline_hotlines_call_suicide
194	39	194_armor_momentary_mennone_crawl
195	39	195_gucci_subliminal_2h_all
196	39	196_anymore_what_do_know
197	39	197_die_want_turds_andni
198	39	198_wellbutrin_paxil_xl_300mg

Bootstrapping Moksha-Erzya Neural Machine Translation from Rule-Based Apertium

Khalid Alnajjar
Rootroo Ltd
khalid@rootroo.com

Mika Hämäläinen
Metropolia University of Applied Sciences
mika.hamalainen@metropolia.fi

Jack Rueter
University of Helsinki
jack.rueter@helsinki.fi

Abstract

Neural Machine Translation (NMT) has made significant strides in breaking down language barriers around the globe. For lesser-resourced languages like Moksha and Erzya, however, the development of robust NMT systems remains a challenge due to the scarcity of parallel corpora. This paper presents a novel approach to address this challenge by leveraging the existing rule-based machine translation system Apertium as a tool for synthetic data generation. We fine-tune NLLB-200 for Moksha-Erzya translation and obtain a BLEU of 0.73 on the Apertium generated data. On real world data, we got an improvement of 0.058 BLEU score over Apertium.

1 Introduction

A significant number of the world's languages are currently at risk of becoming endangered to varying degrees (Moseley, 2010). This endangered status presents particular challenges when it comes to conducting modern NLP research with these languages. The primary issue stems from the fact that many endangered languages lack extensive textual resources that are readily accessible online. Moreover, even when some resources are available, there are concerns regarding the quality of the data, which can be influenced by factors like the author's fluency level, spelling accuracy, and basic character encoding inconsistencies, as discussed in Hämäläinen 2021.

For over two centuries, scholars have been investigating the cohesion and variety within the contemporary Mordvin literary languages, namely Erzya (myv) and Moksha (mdf). The first comprehensive grammatical works on these languages were published in the 1830s, with Moksha in 1838 (Ornatov, 1838) and Erzya in 1838-1839 (Gabelentz, 1839). In the subsequent 180 years, researchers have engaged in extensive fieldwork, compiled grammars, created dictionaries, and worked towards popularizing these languages. Notably, in 2002, the inaugural monolingual Erzya dictionary was published, authored by Abramov (Abramov, 2002), with plans for future expansion. Recent years have also witnessed continued academic interest in the Mordvin languages (Luutonen, 2014; Hamari and Aasmäe, 2015; Kashkin and Nikiforova, 2015; Grünthal, 2016), highlighting their enduring significance in linguistic research.

It is crucial to provide newcomers to language documentation with chances to enhance their comprehension of languages by involving them in projects. A noteworthy period to highlight in this regard is the years spanning from 1988 to 1997 when many of today's researchers were engaged in word processing for the extensive 'Dictionary of Mordvin Dialects', compiled on the basis of language materials whose collection was originated and orchestrated by Prof. Heikki Paasonen at the turn of the twentieth century, and which, when completed, comprised a substantial 2073 pages.

The research conducted in this paper is based on data generated using the rule-based machine translation system Apertium. While rule-based tradition has influenced the current NLP for endangered Uralic languages (cf. Pirinen et al. 2023), our aim is to study the degree to which more modern neural models can be incorporated into the existing paradigm. The largest obstacle in using machine learning models is the scarcity of data available in these languages. Using Apertium to generate training data is our attempt at overcoming this problem.

2 Related work

Many contemporary machine translation models heavily rely on the presence of parallel texts. However, finding parallel texts is a challenging endeavor, particularly when dealing with less-

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 213–218
December 1-3, 2023 ©2023 Association for Computational Linguistics

Erzya input	Moksha output
Лей чиресэ пандыне, */* Пандонть прясо кудыне...	Ляй ширеса пандоня, */* Пандть пряса куданя...
Леесь чуди чипельде пелеве ёнов, лемезэ Ока.	Ляйсь шуди чипельде веньгучка шири, лемезэ Ока.
Сынь каднозь ваномс ды налсо леднемс.	Сань кадондозь ваномс налса ляндендемс.

Table 1: Example of Apertium generated training data

resourced languages across various domains. This challenge becomes even more pronounced when attempting to train data-intensive models for these languages.

A method proposed by Munteanu and Marcu (2005) addresses this issue by utilizing a large, non-parallel but comparable corpus, such as news articles, in conjunction with relatively small parallel corpora from a different domain, like the United Nations corpus. By matching sentences from comparable articles that share the same topic, this method attempts to determine if two sentences are translations of each other. Although this approach enhances translations within the news domain, it's not always viable, especially for extremely low-resource languages like Erzya and Moksha, which lack the necessary comparable corpora.

To develop machine translation models for languages with limited resources, another approach involves leveraging a resource-rich language closely related to the low-resource one as a parent language. This entails acquiring some of the resource-rich language's characteristics, such as syntax and morphology, and transferring them to the low-resource machine translation model (Zoph et al., 2016; Nguyen and Chiang, 2017; Passban et al., 2017; Karakanta et al., 2018). These techniques don't necessarily require parallel texts in the low-resource language but rely on the resource-rich language, which may result in limited coverage of the low-resource language's morphology.

Researchers have also explored methods for constructing parallel texts through crowdsourcing, wherein online workers are tasked with translating expressions into another language (Ambati and Vogel, 2010; Ambati, 2012; Zaidan and Callison-Burch, 2011). Crowdsourcing, however, proves to be a challenging endeavor when dealing with low-resource languages due to the limited number of native speakers. Additionally, the absence of a standardized language form or even linguistic variation further complicates the quality control of crowdsourced translation tasks, even if a significant number of workers are involved.

A different approach proposed by Chahuneau et al. (2013) involves translating English into morphologically complex languages by creating a model that predicts word inflections in the target language. This model is then used to generate synthetic phrases, potentially with new inflections, which are incorporated into the training data alongside a parallel corpus to train a machine translation model.

Hämäläinen and Alnajjar (2019) introduced a method for creating parallel data for low-resource endangered languages with complex morphology. They demonstrated their approach using Finnish as a pilot language, matching the resource limitations to those of Erzya. Additionally, the authors described a technique for automatically aligning the abstract morphosyntactic structures of two languages to generate a set of parallel templates. However, the system could only translate phrases as opposed to complete sentences.

3 Apertium in data generation

Apertium (Forcada et al., 2011) is rule-based machine translation system that is available for several language pairs. There is a special version of the system for endangered languages hosted by GiellaLT (Trosterud, 2017) that has support for Erzya and Moksha. The Erzya-Moksha translation has been developed through a shallow transfer approach (Rueter and Hämäläinen, 2020) and it utilizes FST transducers developed for these languages (Rueter et al., 2020).

We use a monolingual Erzya corpus of around 220 000 sentences (the Erzya-language novel Purgaz (Abramov, 1988)). We feed this corpus to Apertium translator and translate the sentences into Moksha. This way, we will have a parallel corpus of Erzya-Moksha sentences where the target side, namely Erzya is of a high quality and the source side, namely Moksha is synthetically generated using the rule-based system. Apertium is not able to inflect all words, so it tags such words with different tags such as # and @. We remove these extra characters from the data.

Table 1 shows an example of the data we used for training our model. The key notion is that we train

Moksha input	Apertium	Our Model	Gold standard
Минь карматама природать тонафнемонза.	Минь карматано *природать тонавтнемензэ.	Минек карматано природанть тонавтнемензэ.	Минь карматано природанть тонавтнеме.
Природать колга наукати мярьгихть естествознания.	*Природать содалмонтень мерить *естествознания.	Природантьдонть наукантень мерить естествознания.	Природадо наукантень мерить естествознания.
Естествознаниять тейнек пяк оцю значенияц.	*Естествознаниять тенек пек покш *значенияц.	Естествознаниять тенек пек покш значенияц.	Естествознаниянть значениязо миненек пек покш.
Сон лезды лац шарьхкодемс природать.	Сон лезды парсте чарькодемс *природать.	Сон лезды лазтнэнь чарькодиця природанть.	Сон лезды тенек природанть видестэ-парсте чарькодеме.

Table 2: Results on the real world data

our NMT model in an inverse direction. We treat the Moksha output from Apertium as input and the Erzya sentences as output when training the model. This ensures a good and grammatical target representation. This is similar to the back translation methodology described by Sennrich et al. (2016).

The data is split randomly into 80 % training, 10 % validation and 10 % testing. The model is trained only on this Apertium generated dataset.

4 Neural machine translation

We use the NLLB-200 model by Meta (Costa-jussà et al., 2022) to conduct our experiments. The model has been trained to support translation for over 200 languages. Erzya and Moksha are not supported by the model by default. In fact, the only Uralic languages that are supported are Finnish, Estonian and Hungarian - none of which are endangered.

NLLB-200 is a 54.5B parameter Mixture of Experts (MoE) model that has been trained on a dataset containing more than 18 billion sentence pairs. On benchmark evaluations, NLLB-200 outperforms other state-of-the-art models by up to 44%. The model's performance has been validated through extensive evaluations for each of the 200 languages it supports.

We use Transformers Python library (Wolf et al., 2020) to fine-tune the 600M distilled NLLB-200 model[1] for Moksha to Erzya translation. We add two new languages to the tokenizer: mdf_Cyrl and myv_Cyrl for Moksha and Erzya.

We trained the model for 5 epochs. We used weight decay of 0.1 and an initial learning rate of 2e-5. The model was validated after each epoch using BLEU as the validation metric. The final validation BLEU score was 0.85.

5 Results

We have withheld 10% of the data for testing proposes. **Our model achieves a BLEU score of 0.73**. When looking at the results, we can perceive errors coming from the fact that we used Apertium translator such as missing vocabulary resulting in wrong word choice in the translation output as well as minimal transfer rules to address diversity in verbal government and idiomatic expressions. This calls for further comparison of the model with Apertium translations.

We take a relatively small parallel corpus of a natural science book that has been translated into Erzya and Moksha from Russian[2]. The corpus consists of a little over 2000 sentences of human-authored translations. We test both our model and Apertium on this dataset. Neither of the models reaches very good performance when measured by BLEU scores. The Apertium translator gets a BLEU score of 0.037 whereas our model gets a BLEU of 0.095. It is important to note that our model got an improvement of 0.058 BLEU score. The results can be seen in Table 2.

The biggest problem Apertium generated data has is a lack of vocabulary coverage. Our model gets the grammar and morphology correct more frequently than the rule-based Apertium translator because of the good target representation from the high-quality Erzya sentences used in training, although there are instances of multiexponence in morphology as in Природанть+донть 'the nature+about the' which might be explained by incomplete transfer rules in the Apertium translator. However, both Apertium and our model struggle with vocabulary and many of the words are either not translated at all or are translated to wrong words.

However, our result suggest that using Apertium to generate data for an NMT model is a viable way

[1] https://huggingface.co/facebook/nllb-200-distilled-600M

[2] The test example is taken from Russian to Erzya and Moksha translations, see https://urn.fi/urn:nbn:fi:lb-2023042421

of combining the rule-based tradition with latest neural models. In particular, the fact that the NMT model was able to produce better results makes this a worthwhile approach for any endangered language machine translation project. This is even more so in cases where the rule-based translation system has reached a higher level of maturity.

6 Conclusions

In conclusion, our study has revealed significant insights into the performance of our neural machine translation (NMT) model when compared to the rule-based Apertium translator in the context of translation between two endangered languages, Moksha and Erzya.

The primary challenge observed in the Apertium-generated translations was the limited vocabulary coverage. Our NMT model, on the other hand, demonstrated better accuracy in terms of grammar and morphology due to the high-quality Erzya sentences used during training. Nevertheless, both Apertium and our model struggled with vocabulary, resulting in untranslated or incorrectly translated words.

In light of our findings, our results suggest that employing Apertium to generate data for an NMT model represents a viable approach that combines the rule-based translation tradition with state-of-the-art neural models. Notably, our NMT model outperformed Apertium, making this approach valuable for endangered language machine translation projects, particularly when the rule-based translation system has reached a higher level of maturity. This study highlights the potential for leveraging advanced technology to revitalize and preserve endangered languages through more accurate and efficient translation methods.

The findings of this study open up several promising avenues for future research and development in the field of machine translation, particularly for endangered languages. Addressing the challenge of limited vocabulary coverage observed in both the rule-based Apertium translator and the NMT model is critical. Future research could focus on methods to expand and enrich the vocabulary used in training data, possibly through the inclusion of additional linguistic resources or domain-specific terminology.

Augmenting the parallel corpus with more diverse and representative data, including various text genres and dialects, can contribute to better training NMT models. Collecting and curating additional language resources could be a valuable step. Furthermore, we could engage with the community of speakers and experts in Erzya and Moksha languages to gather feedback, improve resources, and establish collaborative efforts in language preservation and machine translation.

These future work directions reflect the ongoing efforts needed to advance the field of machine translation, particularly in the context of preserving and revitalizing endangered languages, and hold the potential to significantly improve the quality and accessibility of translation services for these linguistic communities.

7 Limitations

Our approach requires that there is an existing method of producing translated text in the source language. Furthermore, our approach requires a considerably sized monolingual corpus in the target language. The limitations of the overall system come from limitations in the existing translation system and the monolingual corpus.

The model does not require large computational resources. We trained the model on a desktop PC on an Nvidia RTX3090 GPU. The training was completed in less than a day.

Acknowledgments

This work has been supported by a grant from Oskar Öflunds Stiftelse.

References

Kuzma Abramov. 2002. Валонь ёвтнема валкс. Mordovskoj knizhnoj izdateljstvasj. The manuscript of this dictionary was compiled by the Erzya national writer Kuz'ma Grigorievich Abramov, 1914-2008, whose activities as an Erzya writer spanned nearly 70 years.

Kuzma Abramov. 1988. *Purgaz*. Mordovskoj knižnoj izdatelstvas, Saransk. Online version: https://urn.fi/urn:nbn:fi:lb-2023021601.

Vamshi Ambati. 2012. *Active Learning and Crowdsourcing for Machine Translation in Low Resource Scenarios*. Ph.D. thesis, Pittsburgh, PA, USA. AAI3528171.

Vamshi Ambati and Stephan Vogel. 2010. Can crowds build parallel corpora for machine translation systems? In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 62–65, Stroudsburg, PA, USA. Association for Computational Linguistics.

Victor Chahuneau, Eva Schlinger, Noah A. Smith, and Chris Dyer. 2013. Translating into morphologically rich languages with synthetic phrases. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1677–1687. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Herr Conon von der Gabelentz. 1839. Versuch einer mordwinischen grammatik. In *Zeitschrift für die Kunde des Morgenlandes.*, II. 2–3., pages 235–284, 383–419. Druck und Verlag der Dieterlichschen Buchhandlung., Göttingen.

Riho Grünthal. 2016. *Transitivity in Erzya: Second language speakers in a grammatical focus*, Uralica Helsingiensia, page 291–318. Finno-Ugrian Society, Finland.

Mika Hämäläinen. 2021. Endangered languages are not low-resourced! *Multilingual Facilitation*.

Mika Hämäläinen and Khalid Alnajjar. 2019. A template based approach for training nmt for low-resource uralic languages-a pilot with finnish. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, pages 520–525.

Arja Hamari and Niina Aasmäe. 2015. Negation in erzya. *Negation in Uralic languages*, 108:293.

Alina Karakanta, Jon Dehdari, and Josef van Genabith. 2018. Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, 32(1):167–189.

Egor Kashkin and Sofya Nikiforova. 2015. Verbs of sound in the moksha language: a typological account. *Nyelvtudományi Közlemények*, 111:341–362.

Jorma Luutonen. 2014. Kahden sukupolven ersää – kielenhuoltoa ja muutoksen merkkejä. *Memoires de la Societe Finno-Ougrienne*, 270:187—201.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*, 3rd edition. UNESCO Publishing. Online version: http://www.unesco.org/languages-atlas/.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Comput. Linguist.*, 31(4):477–504.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. *CoRR*, abs/1708.09803.

Pavel Ornatov. 1838. *Mordovskaja grammatika / sostavlennaja na narechij mordvy mokshi Pavlom Ornatovym.* V Sinodalnoj tip., Moskva.

Peyman Passban, Qun Liu, and Andy Way. 2017. Translating low-resource languages by vocabulary adaptation from close counterparts. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 16(4):29:1–29:14.

Flammie Pirinen, Sjur Moshagen, and Katri Hiovain-Asikainen. 2023. GiellaLT — a stable infrastructure for Nordic minority languages and beyond. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 643–649, Tórshavn, Faroe Islands. University of Tartu Library.

Jack Rueter and Mika Hämäläinen. 2020. Prerequisites for shallow-transfer machine translation of mordvin languages: Language documentation with a purpose. In Материалы Международного образовательного салона, pages 18–29. Ижевск: Институт компьютерных исследований.

Jack Rueter, Mika Hämäläinen, and Niko Partanen. 2020. Open-source morphology for endangered mordvinic languages. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 94–100.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

T Trosterud. 2017. Language technology in russia. In ЭЛЕКТРОННАЯ ПИСЬМЕННОСТЬ НАРОДОВ РОССИЙСКОЙ ФЕДЕРАЦИИ: ОПЫТ, ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ, pages 294–298.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Omar F. Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies -*

Volume 1, HLT '11, pages 1220–1229, Stroudsburg, PA, USA. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *CoRR*, abs/1604.02201.

Comparing Transformer and Dictionary-based Sentiment Models for Literary Texts: Hemingway as a Case-study

Yuri Bizzoni
Center for Humanities Computing
Aarhus University, Denmark
yuri.bizzoni@cc.au.dk

Pascale Feldkamp
Center for Humanities Computing
Aarhus University, Denmark
pascale.moreira@cc.au.dk

Abstract

The literary domain continues to pose a challenge for Sentiment Analysis methods, due to its particularly nuanced and layered nature. This paper explores the adequacy of different Sentiment Analysis tools – from dictionary-based approaches to state-of-the-art Transformers – for capturing valence and modelling sentiment arcs. We take Ernest Hemingway's novel *The Old Man and the Sea* as a case study to address challenges inherent to literary language, compare Transformer and rule-based systems' scores with human annotations, and shed light on the complexities of analyzing sentiment in narrative texts. Finally, we emphasize the potential of model ensembles.

1 Introduction

Recent years have seen a significant increase in the methods available for Sentiment Analysis (SA). While dictionary-based approaches like VADER (Hutto and Gilbert, 2014) seem to consistently perform well (Ribeiro et al., 2016), they still struggle when applied to some domains (Elsahar and Gallé, 2019; Ohana et al., 2012; Bowers and Dombrowski, 2021). Transformer-based models provide a much richer semantic representation texts, but also display shortcomings (Tabinda Kokab et al., 2022). While these tools are commonly used to analyze emotive language in contexts like social media (Alantari et al., 2022), their suitability for literary texts remains relatively unexplored. Literary language is particularly intriguing to test SA tools (Chun, 2021), because it often aims to evoke rather than explicitly communicate, operating at multiple narrative levels (Jakobson, 1981; Rosenblatt, 1982; Booth, 1983). In this study, we use *The Old Man and the Sea*, often considered the masterpiece of Ernest Hemingway and exemplary of his philosophy of writing, as a benchmark for testing both rule-based and Transformer-based SA systems.[1] Hemingway's writing style is known for its emotional subtlety, often described as an "iceberg" or "omissive" writing, that evokes more than it describes: "the emotion is plentiful, though hidden but not exposed" (Daoshan and Shuo, 2014). With its directness and limited use of figurative language (Heaton, 1970), Hemingway avoids "overt emotional display" (Strychacz, 2002) in a way that may pose a particular challenge to SA. Building on the literary analysis tradition that seeks to model sentiment arcs in literary texts (Jockers, 2014; Maharjan et al., 2018; Elkins, 2022), we apply various methods for sentiment annotation to the sentences of the novel and compare them to a benchmark of human annotations.

2 Related works

In literary studies, what is often called the "affective turn" (Armstrong, 2014) has led to a stronger focus on sentiment expressed in narrative texts (Ngai, 2007), and SA has often been employed in computational literary studies to profile texts and model the "shape of stories" (Reagan et al., 2016a). To capture meaningful aspects of the reading experience, previous works tested the potential of SA (Alm, 2008; Jain et al., 2017) at the word (Mohammad, 2011, 2018), sentence (Mäntylä et al., 2018), or paragraph level (Li et al., 2019) to model narrative arcs (Kim and Klinger, 2018; Reagan et al., 2016a; Jockers, 2014). Sentiment arcs have been used to evaluate literary texts in terms of shape or plot (Reagan et al., 2016a) progression (Hu et al., 2020), and mood (Öhman and Rossi, 2022). Certain shapes or arc dynamics have been connected to reader appreciation, considering both simple and more complex narratives (Bizzoni et al., 2022a, 2023), and Bizzoni et al. (2023) have shown that sentiment features, such as measures

[1] Link to the annotated text (human and automatic annotations): https://github.com/PascaleFMoreira/Annotated_Hemingway

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 219–228
December 1-3, 2023 ©2023 Association for Computational Linguistics

of sentiment arc progression, have an effect even compared to the predominantly stylistic features usually employed for this type of task (Koolen et al., 2020; Maharjan et al., 2017). As such, modelling sentiment arcs holds potential for gaining a more in-depth understanding of how narratives, in their unfolding, affect readers. However, both the validity of the dictionary-based approaches and the adequacy of methods for detrending arcs (Gao et al., 2016) have been controversial in literary SA (Swafford, 2015; Hammond, 2017; Elkins, 2022; Rebora, 2023). For example, dictionary-based methods seem to perform well even on so-called "non-linear" narratives (Richardson, 2000; Elkins and Chun, 2019) although they appear to do poorly on a word-basis (Reagan et al., 2016b). On the other hand, more recent Transformer-based approaches have shown both potential and pitfalls in the analysis of sentiment (Elkins, 2022).

3 Methods

3.1 Human Annotation

The first contribution of this paper is to provide a valence-annotated version of *The Old Man and the Sea*. Human annotators (n=2) read it from beginning to end and scored its 1923 sentences on a 1 to 10 valence scale: 1 signifying the lowest, and 10 the highest valence. Here, valence was intended as the sentiment expressed by the sentence. The annotators were instructed to avoid rating how a sentence made them feel and to try to report only on the sentiments actually embedded in the sentence, i.e., to think about the valence of each sentence individually, without overthinking the story's narrative to reduce contextual interpretation. This naturally is far from an obvious or objective task, which created several interesting cases of uncertainty or ambiguity.

Both annotators have extensive experience of literary analysis, and hold degrees in literature.[2] They worked independently, not discussing nor subsequently changing scores. The task was not explicitly categorical: the annotators could use in principle decimals or even more fine-grained representations of their perceived valence. Nonetheless, both annotators resorted to using discrete values only. As mentioned, *The Old Man and the Sea* is an advantageous case-study for SA. While the story arc is linear and the style is simple, it is often ambivalent, shifting perspectives and narrative sympathies between the natural and human world, so that it can be difficult to annotate even for a human reader. For example, the sentence "Then the fish came alive, with his death in him, and rose high out of the water showing all his great length and width and all his power and his beauty" is stylistically simple, but offers a tension between contrasting emotions that challenges linear valence scales.

Accordingly, the correlation between the human annotators is not perfect, albeit very robust (Pearson: 0.652; Spearman: 0.624). The Cohen-Kappa score is 0.342. While this is relatively low, seeing as the annotators were working on a continuous valence space which was discrete in ten categories, we consider correlation measures to be more adequate than categorical inter-annotator agreement measures. A representation of the detrended sentiment arc of each annotator is visualized in the Appendix, along with their detrended mean.

After detrending the arcs, the correlation between the annotators' arcs is much more robust, with a Pearson correlation of 0.92. In short, this means that humans differ more on their sentence-by-sentence judgment of valence than they differ on the overall sentiment arc of the novel. Detrended arcs are in fact an attempt to draw the shape of the overall sentiment progress of a text, independently from the "noise" of individual sentences' ups-and-downs. As such, they tend to be more linear, more robust, and to elicit higher correlations between models.

3.2 Automatic Annotation

All annotations were performed on a sentence-basis (not considering context).[3]

3.2.1 Transformers

For the automatic annotation of the novel's sentences we used four SOTA Transformers: (i) DistilBERT base uncased, fine-tuned on SST2 (Sanh et al., 2020), (ii) BERT base uncased, fine-tuned on product reviews for SA (Peirsman, 2020), (iii) roBERTa base, fine-tuned for SA on tweets (Barbieri et al., 2020), (iv) roBERTa base, fine-tuned for multilingual SA on tweets (Barbieri et al., 2022).[4]

[2] Both were academics, male and female, at ages 31 and 34, who were non-native but very proficient English speakers, and who finished their literature degree (MA and BA) finished 1, respectively 12 years ago (the BA).

[3] Sentences were tokenized using the nltk tokenize package: https://www.nltk.org/api/nltk.tokenize.html

[4] We included the multilingual roBERTA to test this model for future work on multilingual literary corpora.

The first model returns two possible categories, *positive* or *negative*; models 3 and 4 also have the *neutral* category. Instead, model 2 returns five different categories, from 1, most negative, to 5, most positive. It's important to remember that unlike dictionary-based models, Transformers' output is categorical in nature. To use their output for representing continuous sentiment arcs, we have used the confidence score of their labels as a proxy for sentiment intensity. So if the model classifies a sentences as *positive* with a confidence of, for example, 0.89, we interpret it as a valence score on the sentence of +0.89. If the model classifies a sentences as *negative* with a confidence of 0.89, we interpret it as a valence score on of the sentence of -0.89. However, we couldn't do the same for the *neutral* category (or category 3 in system (iii)), so we simply converted these cases to a score of 0. Naturally this may make the comparison less fair for these models than for the models already designed for a continuous scoring approach. On the other hand, our quest is precisely to find out, which model(s) approximate a human continuous valence rating on literary texts.

3.2.2 Dictionary-based models

To compare against Transformers, we chose two dictionary-based approaches: (i) the nltk implementation of VADER (Hutto and Gilbert, 2014), arguably the most widespread dictionary-based method for SA. (ii) Syuzhet (Jockers, 2014), a widespread implementation, designed to model literary arcs. The dictionary is extracted from 165,000 human coded sentences from contemporary literary novels, developed in the Nebraska Literary Lab (Jockers, 2015b). Both models dictionary- and rule-based, and return continuous scores ranging from -1 (negative) to +1 (positive).

3.3 Detrending sentiment arcs

A sentiment arc refers to a simple 1d representation of sections of a literary work (e.g., the valence of words, sentences or paragraphs). Because narratives and derived arcs based on the valences are inherently noisy and nonlinear, studies typically apply some technique for detrending or "smoothing" arcs to reduce noise and extract the global narrative trends - from a simple moving average window to more complex noise reduction techniques (Chun, 2021; Jockers, 2015a; Bizzoni et al., 2021; Gao et al., 2016). As wavelet approaches typically used for noise reduction are not ideal for nonlinear series, Jianbo Gao et al. (2010) proposed an adaptive filtering technique for nonlinear series. Studies have demonstrated the usefulness of adaptive filtering applied to sentiment arcs, especially in the context of estimating dynamics of sentiment arcs (Hu et al., 2020; Bizzoni et al., 2022b). Arcs are based on the second polynomial fit (m=2).

4 Results

To evaluate the models we use the average of the annotators' scores. In Table 1 we present the correlations between each model and the human baseline. We also add the correlations with two "ensemble" approaches: the average of all SA models' outputs, and a select average of the outputs of only Roberta, Roberta xlm and Syuzhet: the three best performing models.

Our results show that large pretrained Transformers correlate with human judgments on the valence of sentences better than the rule-based VADER and Syuzhet. Thus, despite Transformer's output on each sentence being categorical, it appears that their confidence scores can be successfully used as proxies for valence intensity even on literary sentences (see the Appendix for a detailed plot of raw values). Still, it is notable that the dictionary-based systems outperform half our Transformer population. Interestingly, the correlation of each model with each individual human is *lower* than the correlation of each model with the average human annotation (Table 1) - in other words, sentiment seems to act almost as an objective measure, with individual stochastic "errors" reduced through repeated annotation. If we observe the sentences with the highest disagreement between (average) human judgment and the best performing Transformer, Roberta XLM, we find that these sentences tend to be short, where the model displays a negativity bias; while the sentences where the best performing rule-based model, Syuzhet, is most removed from the human evaluation appear to be long sentences with complex semantic interplays, for which it displays a positivity bias. Finally, the sentences with most disagreement between the two models are often sentences that were also difficult for human annotators. In the Appendix we show a small selection of such sentences (Table 4).

When detrending the series of valences, we find that the picture changes: Syuzhet outperforms all Tranformers (Table 1). It is possible that in the case of Syuzhet, errors at the level of raw scores,

	DistilBert	Bert	Roberta	Roberta_xlm	Vader	Syuzhet	Average	Select
Kendall τ	0.39	0.28	**0.50**	**0.50**	0.36	0.36	0.48	0.50
Spearman r	0.51	0.36	0.57	**0.59**	0.43	0.45	0.59	0.61
Pearson r	0.42	0.36	**0.63**	**0.63**	0.46	0.48	0.65	0.70
Pearson r**, per annot.**	.41/.48	.35/.30	.59/.56	.59/.55	.45/.39	.46/.41	.62/.55	.66/.61
Kendall τ	0.62	0.49	0.75	0.73	0.41	**0.84**	0.83	0.84
Spearman r	0.80	0.68	0.90	0.89	0.57	**0.96**	0.96	0.96
Pearson r	0.80	0.71	0.90	0.85	0.68	**0.96**	0.96	0.96
Pearson r**, per annot.**	.88/.71	.70/.69	.92/.85	.87/.81	.62/.70	.90/.97	.96/.93	.95/.93

Table 1: **Top**: correlations between *raw* annotations and the human mean values. The last row indicates the Pearson correlation per method to each annotator individually. **Bottom**: Correlations between *detrended* annotations and the human mean values. For all correlations, *p-values* < 0.01.

where humans set a negative and Syuzhet a positive score (see Appendix, Fig. 4),[5] are big enough to impact the overall correlation with human annotations, but are still few enough to be "cancelled" out in detrending, so that dictionary-based arcs are the closest to the human arc. The detrending essentially flattens out raw scores, so that scores that are proximate are more alike. In this sense, detrending gives us a pictures of the annotation tendencies at each point of the arc, and smoothens out scores that diverge suddenly from the tendencies.

5 Analysis

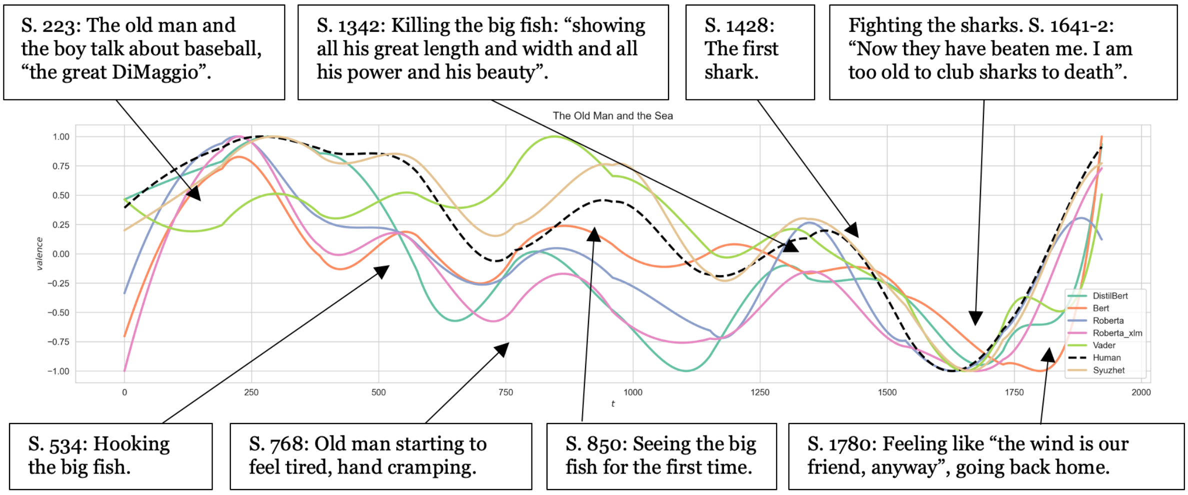
Literary language is a challenge to SA due to its subtlety and complexity. Narrative sentences can be as complex as those of any other domain, yet because literary texts aim for their readers to experience rather than just be informed, they seem specially difficult to annotate. Looking at the human scores of *The Old Man and the Sea*, we found that annotators used almost the whole range (1 to 10), going from 2 to 9. Though annotators were instructed not to overthink the narrative to reduce contextual scoring, this was not always easy. Hemingway's direct style partly facilitated annotation, e.g.: (*"Fish," he said, "I love you and respect you very much"*), but underlying complexity sometimes sparked uncertainty and disagreement for human annotators. Despite being negative agents in the story, the sharks, for example, are still described as "beautiful", and the protagonist is portrayed as both "beat" and "undefeated". Several of the larger inter-annotator disagreements were often due to the presence of co-existing valences in the same sentence. Several of such sentences elicited differing judgments from the models as well: for example the sentence "*The old man hit him on the head for kindness and kicked him, his body still shuddering, under the shade of the stern*" elicited scores of 6 and 2 from the annotators, -.97 from DistilBert and +.46 from VADER (normalized values).

We have already observed that almost all models correlate less with individual annotators than with the mean of the two annotators, an effect that is magnified when we also compute the mean of all the models' scores: the average annotation of all the models (after normalization) correlates with the human judgments better or as well as the individual models, both for the raw scores and for the detrended arcs.

6 Discussion and Conclusions

For this case-study in comparing sentiment annotation methods for literary analysis, we have compared the correlations between human annotations and several SA systems' annotations of the sentences of the novel *The Old Man and the Sea*. While sentiment analysis is often tackled as a classification problem (with two or three categories at most), we found this approach to be exceedingly coarse-grained to verify the efficacy of SA models on literary texts, and we preferred to model it as a continuous scoring task. Most of the time human annotators would have been unable to fit a sentence into a binary classification, and the most interesting behaviours of the models happen when looking at their ability to position a sentence on a nuanced continuum. Naturally, it is now possible to operate the opposite operation and convert the continuous annotations into two or three categories, to compare them directly with the Transformer's outputs.

[5] This may be due to systematic errors, such as the issue with negations in Syuzhet.



S. 223: The old man and the boy talk about baseball, “the great DiMaggio”.

S. 1342: Killing the big fish: “showing all his great length and width and all his power and his beauty”.

S. 1428: The first shark.

Fighting the sharks. S. 1641-2: “Now they have beaten me. I am too old to club sharks to death”.

S. 534: Hooking the big fish.

S. 768: Old man starting to feel tired, hand cramping.

S. 850: Seeing the big fish for the first time.

S. 1780: Feeling like “the wind is our friend, anyway”, going back home.

Figure 1: Arcs of *The Old Man and the Sea* based on various methods, with manual annotations of narrative events. The added dashed line represents the mean value of human annotators.

	DistilBert	Bert	Roberta	Roberta_xlm	Vader	Syuzhet
Avg. difference	0.86	0.48	0.19	0.26	0.23	0.16
Std	0.22	0.32	0.22	0.26	0.24	0.15

Table 2: Mean difference and standard deviation between human and model valence.

We have observed interesting differences between Transformer- and dictionary-based methods. Still, it should be noted that our analysis was performed on one story only, even though the particular example of *The Old Man and the Sea* appears particularly apt as case-study for Sentiment Analysis, considering its emotionally understating literary style. Despite being categorical in nature, the largest Transformers of our collection proved to hold strong correlations with human judgments in the sentence-level annotation – higher than the dictionary-based VADER and Syuzhet. When looking at the detrended version of the arcs, the picture is reversed: despite serious shortcomings of the tool (Kim, 2022), detrended arcs made with the Syuzhet package appear to be the most closely related to the detrended version of human arcs (Fig. 1). In both cases, the best results are achieved when using both Transformer and dictionary-based systems, as they appear to be at least partly complementary, and our best model correlates with the mean human score almost as much as humans correlated with each other (Table 1). We have observed that average human judgments seem to be more aligned to models than individual judgments, and average automatic scores from different sources seem to work better than the scores of any individual model. Moreover, at the sentence level, while roBERTa correlated with human judgments best, VADER and Syuzet are closer to the human intensities: on average, VADER and Syuzhet have a smaller mean distance from human intensity (as does the roBERTa), and a lower standard deviation (Table 2). [6] Beyond providing the best correlation with human judgments, it's possible that a compound approach, integrating the scores of two or more models, would be greatly beneficial for something else: the detection of confounding or polarizing sentences, likely to elicit opposite scores. Some of the sentences with the largest difference between rule-based and Transformer-based scores are beautifully complex to judge for human readers alike, such as the sentence that elicited the the highest disagreement between models: *“I killed him in self-defense," the old man said aloud. “And I killed him well."*

Limitations

As sentiment annotation is a difficult task, this study has attempted to make the process as robust as possible, and we have sought to make our

[6] We also observe that, when inspecting raw scores, Transformers seem to be more “extreme" in their judgement than human and dictionary-based models. See Appendix for a visualization.

procedure by various SA methods as transparent as possible. Regardless, identifying sentiment in text is always subjective and difficult to measure, and may be subject to cultural understandings of sentiment expression – which inevitably situates our analysis in the Anglophone cultural context. Moreover, it should be noted that our annotators were academics, and though their annotation may reflect their knowledge of literary devices and language, it also reflects the cultural understandings of a particular class. As a case-study moving towards a comparison and better understanding of sentiment analysis methods, it should also be noted that the analysis limits itself to one, and a particularly canonic, Anglophone literary novel. We trust that any interpretation of our findings will have these limitations in mind.

References

Huwail J. Alantari, Imran S. Currim, Yiting Deng, and Sameer Singh. 2022. An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews. *International Journal of Research in Marketing*, 39(1):1–19.

Ebba Cecilia Ovesdotter Alm. 2008. *Affect in* text and speech*. Phd thesis, University of Illinois at Urbana-Champaign.

Nancy Armstrong. 2014. The Affective Turn in Contemporary Fiction. *Contemporary Literature*, 55(3):441–465. Publisher: [Board of Regents of the University of Wisconsin System, University of Wisconsin Press].

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification.

Yuri Bizzoni, Pascale Moreira, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2023. Sentimental matters - predicting literary quality by sentiment analysis and stylometric features. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 11–18, Toronto, Canada. Association for Computational Linguistics.

Yuri Bizzoni, Telma Peura, Kristoffer Nielbo, and Mads Thomsen. 2022a. Fractal sentiments and fairy tales-fractal scaling of narrative arcs as predictor of the perceived quality of Andersen's fairy tales. *Journal of Data Mining & Digital Humanities*, NLP4DH.

Yuri Bizzoni, Telma Peura, Kristoffer Nielbo, and Mads Thomsen. 2022b. Fractality of sentiment arcs for literary quality assessment: The case of nobel laureates. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 31–41, Taipei, Taiwan. Association for Computational Linguistics.

Yuri Bizzoni, Telma Peura, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2021. Sentiment dynamics of success: Fractal scaling of story arcs predicts reader preferences. In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 1–6, NIT Silchar, India. NLP Association of India (NLPAI).

Wayne C. Booth. 1983. *The Rhetoric of Fiction*, 2nd edition edition. University of Chicago Press, Chicago.

Katherine Bowers and Quinn Dombrowski. 2021. Katia and the Sentiment Snobs. Blog: Datasitter's Club.

Jon Chun. 2021. SentimentArcs: A Novel Method for Self-Supervised Sentiment Analysis of Time Series Shows SOTA Transformers Can Struggle Finding Narrative Arcs. ArXiv:2110.09454 [cs].

MA Daoshan and Zhang Shuo. 2014. A discourse study of the Iceberg Principle in *A Farewell to Arms*. *Studies in Literature and Language*, 8(1):80–84.

Katherine Elkins. 2022. *The Shapes of Stories: Sentiment Analysis for Narrative*. Cambridge University Press.

Katherine Elkins and Jon Chun. 2019. Can Sentiment Analysis Reveal Structure in a Plotless Novel? ArXiv:1910.01441 [cs].

Hady Elsahar and Matthias Gallé. 2019. To Annotate or Not? Predicting Performance Drop under Domain Shift. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2163–2173, Hong Kong, China. Association for Computational Linguistics.

Jianbo Gao, Matthew L Jockers, John Laudun, and Timothy Tangherlini. 2016. A multiscale theory for the dynamical evolution of sentiment in novels. In *2016 International Conference on Behavioral, Economic and Socio-cultural Computing (BESC)*, pages 1–4. IEEE.

Adam Hammond. 2017. The double bind of validation: distant reading and the digital humanities' "trough of disillusionment". *Literature Compass*, 14(8):e12402. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/lic3.12402.

C. P. Heaton. 1970. Style in *The Old Man and the Sea*. *Style*, 4(1):11–27. Publisher: Penn State University Press.

Qiyue Hu, Bin Liu, Mads Rosendahl Thomsen, Jianbo Gao, and Kristoffer L Nielbo. 2020. Dynamic evolution of sentiments in Never Let Me Go: Insights from multifractal theory and its implications for literary analysis. *Digital Scholarship in the Humanities*, 36(2):322–332.

Clayton Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Swapnil Jain, Shrikant Malviya, Rohit Mishra, and Uma Shanker Tiwary. 2017. Sentiment analysis: An empirical comparative study of various machine learning approaches. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 112–121, Kolkata, India. NLP Association of India.

Roman Jakobson. 1981. Linguistics and poetics. In *Linguistics and Poetics*, pages 18–51. De Gruyter Mouton.

Jianbo Gao, H. Sultan, Jing Hu, and Wen-Wen Tung. 2010. Denoising Nonlinear Time Series by Adaptive Filtering and Wavelet Shrinkage: A Comparison. *IEEE Signal Processing Letters*, 17(3):237–240.

Matthew Jockers. 2014. A Novel Method for Detecting Plot. Matthew L. Jockers Blog.

Matthew Jockers. 2015a. Revealing Sentiment and Plot Arcs with the Syuzhet Package. Matthew L. Jockers Blog.

Matthew L. Jockers. 2015b. *Syuzhet: Extract Sentiment and Plot Arcs from Text*.

Evgeny Kim and Roman Klinger. 2018. A survey on sentiment and emotion analysis for computational literary studies. *arXiv preprint arXiv:1808.03137*.

Hoyeol Kim. 2022. Sentiment analysis: Limits and progress of the Syuzhet package and its lexicons. *Digital Humanities Quarterly*, 16(2).

Corina Koolen, Karina van Dalen-Oskam, Andreas van Cranenburgh, and Erica Nagelhout. 2020. Literary quality in the eye of the Dutch reader: The national reader survey. *Poetics*, 79:1–13.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting BERT for end-to-end aspect-based sentiment analysis. pages 34–41.

Suraj Maharjan, John Arevalo, Manuel Montes, Fabio A. González, and Thamar Solorio. 2017. A multi-task approach to predict likability of books. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1217–1227, Valencia, Spain. Association for Computational Linguistics.

Suraj Maharjan, Sudipta Kar, Manuel Montes, Fabio A. González, and Thamar Solorio. 2018. Letting Emotions Flow: Success Prediction by Modeling the Flow of Emotions in Books. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 259–265, New Orleans, Louisiana. Association for Computational Linguistics.

Saif Mohammad. 2011. From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114, Portland, OR, USA. Association for Computational Linguistics.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Mika V. Mäntylä, Daniel Graziotin, and Miikka Kuutila. 2018. The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. 27:16–32.

Sianne Ngai. 2007. *Ugly Feelings*. Harvard University Press, Cambridge, MA.

Bruno Ohana, Sarah Jane Delany, and Brendan Tierney. 2012. A Case-Based Approach to Cross Domain Sentiment Classification. In *Case-Based Reasoning Research and Development*, Lecture Notes in Computer Science, pages 284–296, Berlin, Heidelberg. Springer.

Emily Öhman and Riikka H. Rossi. 2022. Computational exploration of the origin of mood in literary texts. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 8–14, Taipei, Taiwan. Association for Computational Linguistics.

Yves Peirsman. 2020. nlptown/bert-base-multilingual-uncased-sentiment · Hugging Face.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016a. The Emotional Arcs of Stories Are Dominated by Six Basic Shapes. *EPJ Data Science*, 5(1):1–12.

Andrew J. Reagan, Brian Tivnan, Jake Ryland Williams, Christopher M. Danforth, and Peter Sheridan Dodds. 2016b. Benchmarking sentiment analysis methods for large-scale texts: A case for using continuum-scored words and word shift graphs. (arXiv:1512.00531). ArXiv.

Simone Rebora. 2023. Sentiment Analysis in Literary Studies. A Critical Survey. *Digital Humanities Quarterly*, 17(2).

Filipe N. Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1–29.

Brian Richardson. 2000. Linearity and Its Discontents: Rethinking Narrative Form and Ideological Valence. *College English*, 62(6):685–695.

Louise M. Rosenblatt. 1982. The literary transaction: Evocation and response. *Theory Into Practice*, 21(4):268–277.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv:1910.01108 [cs].

Thomas Strychacz. 2002. "The sort of thing you should not admit": Ernest Hemingway's aesthetic of emotional restraint. In Milette Shamir and Jennifer Travis, editors, *Boys Don't Cry? Rethinking Narratives of Masculinity and Emotion in the U.S.*, pages 141–166. Columbia University Press.

Annie Swafford. 2015. Problems with the Syuzhet Package. *Anglophile in Academia: Annie Swafford's Blog*.

Sayyida Tabinda Kokab, Sohail Asghar, and Shehneela Naz. 2022. Transformer-based deep learning models for the sentiment analysis of social media data. *Array*, 14:100157.

A Appendix

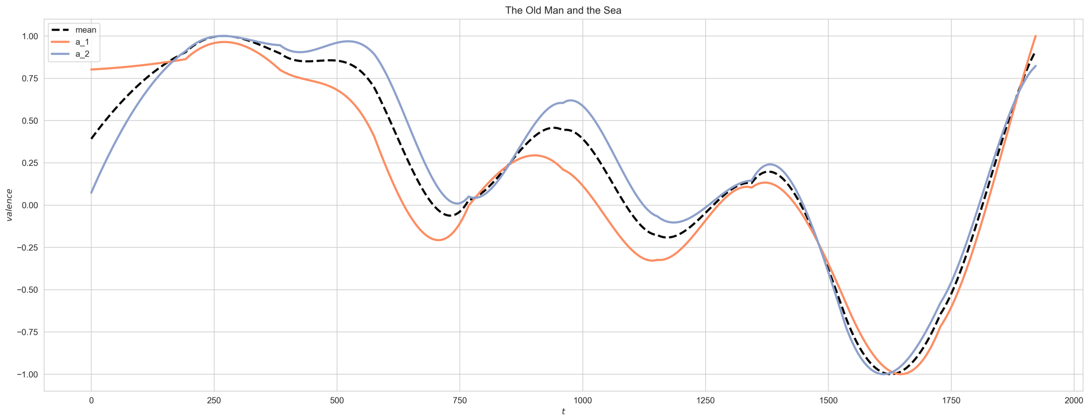


Figure 2: Arc of *The Old Man and the Sea* based on annotator (n=2) values. The dashed line represents the mean value of annotators.

	DistilBert	**Bert**	**Roberta**	**Roberta_xlm**	**Vader**	**Syuzhet**
Raw	0.13	0.11	0.39	0.33	0.15	-1.03
Detrended	0.34	-0.38	0.43	-0.11	0.23	0.91

Table 3: R2 scores for time series compared to the human mean values.

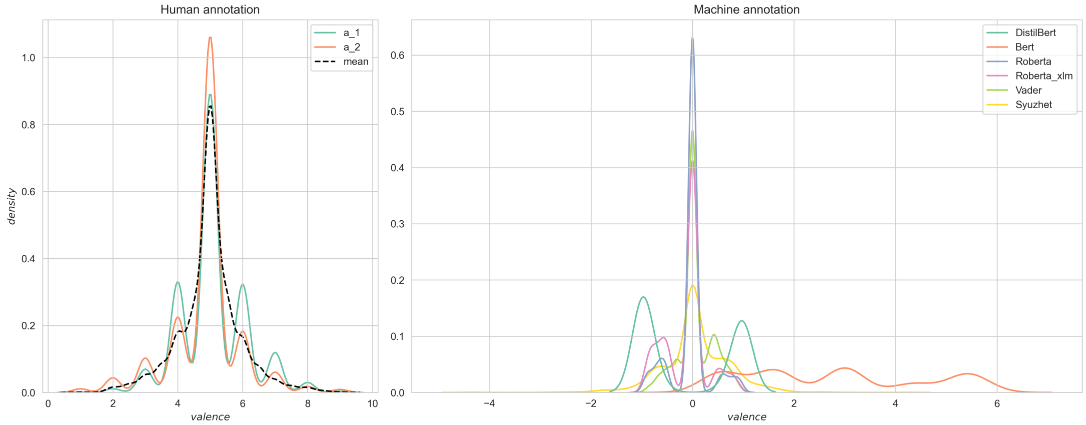


Figure 3: Kernel density plots visualize the distributions of values (0 or neutral being the most common). Note that value ranges differ: the BERT model, for example, assigns valence on a 5-point scale, while human annotators could assign any (round) value between 0 and 10.

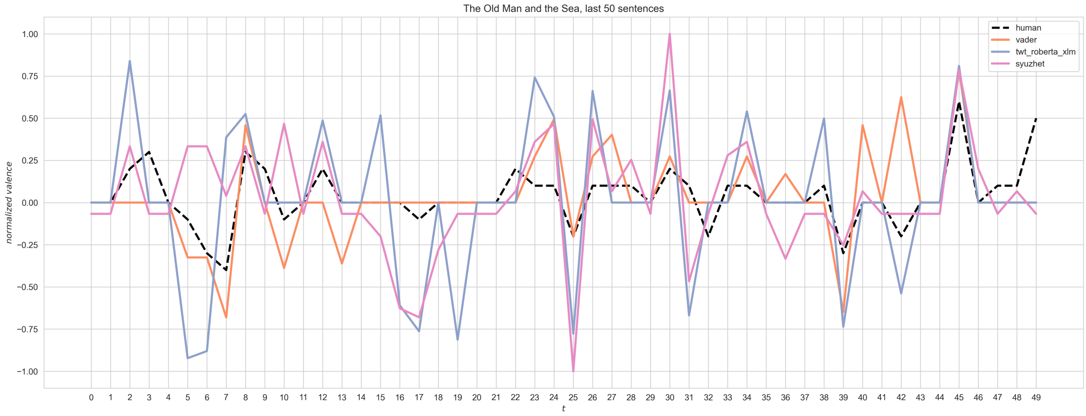


Figure 4: Arc of the last 50 sentences of *The Old Man and the Sea* with on transformer and dictionary-based annotation. The added dashed line represents the mean value of human annotators. Note that sentences like [5]: “I am not lucky" and [10] “I do not care" are systematically misjudged as positive in the Syuzhet annotation despite the negations.

Sentence	Roberta_xlm	Syuzhet	Human
They were immune to its poison	-.87	-.05	.3
Perhaps he is too wise to jump	-.68	-.14	.3
"I wish the boy was here," he said aloud and settled himself against the rounded planks of the bow and felt the strength of the great fish through the line he held across his shoulders moving steadily toward whatever he had chosen.	.42	.63	-.1
There is no one worthy of eating him from the manner of his behaviour and his great dignity.	-.92	.2	-.1
The old man’s head was clear and good now and he was full of resolution but he had little hope	-.85	.15	-.2

Table 4: Examples of sentences with the largest disagreement *between machine and (normalized) human score* for Roberta XLM (upper rows of the table) and Syuzhet (central rows the table). Roberta XLM is most off track for short, relatively ambiguous sentences; Syuzhet appears to disagree more with long and complex sentences. Examples of sentences that instead elicit a large disagreement *between the two models* are in the lower rows the table. These sentences are often also complex for human annotators to judge.

Sentence	DistilBert	Bert	Roberta	Roberta_xlm	Vader	Syuzhet	Human
Then he felt the gentle touch on the line and he was happy.	**0.9998**	4.42	0.94	0.68	0.76	0.42	6.5
Blessed art thou among women and blessed is the fruit of thy womb, Jesus.	0.9982	**5.91**	0.84	0.86	0.83	0.45	6.5
"Tomorrow is going to be a good day with this current," he said.	0.9991	4.37	**0.98**	0.89	0.44	0.19	6.5
Bed will be a great thing.	0.9996	5.59	0.95	**0.91**	0.62	0.14	7.5
But he was such a calm, strong fish and he seemed so fearless and so confident.	0.9997	5.38	0.85	0.75	**0.95**	0.72	**8.0**
The boy had given him two fresh small tunas, or albacores, which hung on the two deepest lines like plummets and, on the others, he had a big blue runner and a yellow jack that had been used before; but they were in good condition still and had the excellent sardines to give them scent and attractiveness.	0.9972	4.5	0.8	0.45	0.94	**1.0**	7.0

Table 5: To give a short overview of the models’ comparative performance, we present the sentences that elicited the highest score for each model.

Study on the Domain Adaption of Korean Speech Act using Daily Conversation Dataset and Petition Corpus

Youngsook Song
Sionic AI Inc.
Seoul, Korea
song@sionic.ai

Won Ik Cho
Seoul National University[*]
Seoul, Korea
tsatsuki@snu.ac.kr

Abstract

In Korean, quantitative speech act studies have usually been conducted on single utterances with unspecified sources. In this study, we annotate sentences from the National Institute of Korean Language's Messenger Corpus and the National Petition Corpus, as well as example sentences from an academic paper on contemporary Korean vlogging, and check the discrepancy between human annotation and model prediction. In particular, for sentences with differences in locutionary and illocutionary forces, we analyze the causes of errors to see if stylistic features used in a particular domain affect the correct inference of speech act. Through this, we see the necessity to build and analyze a balanced corpus in various text domains, taking into account cases with different usage roles, e.g., messenger conversations belonging to private conversations and petition corpus/vlogging script that have an unspecified audience.

1 Introduction

People use statements to reveal the intent of a proposition or to express their promises or emotions. However, similar can be applied to questions. Generally speaking, interrogatives are uttered in situations where the speaker does not know the relevant information but assumes that the listener does. To express a question, a speaker would use an interrogative ending and a question mark in written language, or a rising intonation in spoken language. Nonetheless, the use of interrogative endings, question marks, or rising intonation does not necessarily constitute interrogative speech. In this regard, the examples given by Song (2010) and Park (2019) are as follows.

(1) a. Mr. Lee: 바보.. 메주야 넌! (*Fool.. you idiot!*)
Bom: 아휴! 내가 왜 메주야! (*Ahhh! Why am I an idiot!*) (Song (2010): 98)

b. 나라의 운명을 외국의 손에 맡겨서야 되겠습니까 (*Do we hand over the fate of our country to foreigners?*) / 이런 걸 누가 먹겠습니까 (*Who would eat something like this*) (Park (2019): 16)

Example (1a) emphasizes the speaker's negative emotions by utilizing a distinctive speech style, particularly through the use of the interrogative ’*why*’ by the speaker in the ’Bom’ example. (1b) Despite adopting the forms of Yes-No-Questions and Wh-Questions, it is not readily classified as an interrogative speech act because it is used to emphasize the opinion rather than to elicit information. Notably, humans tend to adeptly comprehend the speaker's intention, even when a disparity exists between explicit form and implicit intent. However, artificial intelligence (AI) models may face challenges in such interpretive tasks. Consequently, as exemplified above, speech act annotations could contribute to enhancing the utterance performance of AI models, particularly in instances where the latent meaning of an utterance diverges from its manifest content.

In this study on the Korean speech act, an attempt is made to measure the performance of AI models analyzing distinctions between locutionary and illocutionary force, especially when disparities exist between the two. For the purpose of performance measurement, frequently mispredicted speech acts are typified. For instance, there are cases, such as example (1b), where the emphasis on intention may be misinterpreted as a question because the context is not specified. This is similar to how, without context, it is difficult for humans to categorize ‘speech act’ into specific categories. In circumstances where distinguishing speech act is possible only if given context, the likelihood of models correctly identifying the answer may become notably low in the case of sentence-level annotated data. Conversely, even without any con-

[*]Work done after graduation.

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 229–234
December 1-3, 2023 ©2023 Association for Computational Linguistics

Statement	*Statement*	Declarative utterances that include or convey proposition
	Future Intention	Utterances that describe the speaker's will or promises
	Sarcasm/Humor	Utterances that convey the speaker's sarcasm or humor towards the object
Suggestion	*Suggestion*	Commands or requests, including short directions
Exclamation	*Exclamation*	Utterances with expressions that display daily emotions
Question	*Yes-No-Question*	Polar and multiple choice questions
	Wh-Question	Open questions that require further answers
	Rhetoric-question	Questions that do not require an answer from the addressee
Greeting	*Greeting*	Conventional greetings including optatives
	Adress term	Addressing others with name or title

Table 1: Speech act annotation criteria.

text provided, if a specific speech act is commonly utilized in a particular discourse situation, anticipations of relatively effortless performance improvements can be posited through the construction of a sufficiently large and diverse corpus. Therefore, this study intends to scrutinize, in detail, various instances such as National Institute of Korean Language (NIKL)'s Messenger Corpus (2022) (which was updated from 2020 NIKL corpora (NIKL, 2020)), excerpts from an academic paper on contemporary Korean vlogging, and the titles of public petitions (those are in oratory style), to identify under which circumstances models incur errors in speech act classification. Initially, after annotating speech acts in conversations within the Messenger corpus, we undergo automatic classification with a widely used pretrained language model (PLM), the bert-base-multilingual-cased model (Devlin et al., 2018).

2 Speech Act Annotation

2.1 Speech Act Theory

Regarding the definition of speech acts, this study adopts Austin and Searle's speech act theories. Austin (1962) categorizes speech acts into Commissives, Verdictives, Exercitives, Behabitives, and Expositives, and describes the speaker's 'utterance intention' as an illocutionary force, and they more adapted in Searle (1976) to a criteria that is widely applicable. Though Stolcke et al. (2000) added rhetorical question as a notable dialogue act among other forty speech act classes, in a more recent and systematic approach, Bunt et al. (2010) encompassed the tripartite classification of questions, namely propositional questions, check questions, set questions/choice questions. In a relatively recent study on Korean, Cho and Kim (2022) distinguished between usual questions and rhetorical questions (denoted as RQ) within the questions, and also within commands; they categorized directives as commands if they solicited a specific action, and otherwise as rhetorical commands (denoted as RC), which is particularly significant in optatives.

In this study, we also deem it necessary to distinguish between the locutionary act, which pertains to the sentence's meaning and directive action, and the illocutionary act, which involves subsequent speech actions such as promises, commands, and coercions. Furthermore, if a developed model can comprehend and generate speech acts based on these distinctions, it could be applied to and utilized across various industrial domains.

2.2 Data Annotation

For the annotation of data that is adopted in the model training, NIKL Messenger Corpus (NIKL, 2020) was utilized by collecting a total of 33,138 sentences from 3,840 files. The source data was collected from free conversation of the participants and is available under application from NIKL online page[1].

In addition to the Messenger Corpus, a more challenging evaluation set with 125 sentences was constructed by extracting examples from a research paper on contemporary Korean vlogging and microblogging (Park, 2022) and bringing titles from public petitions[2]. They are known to characteristically reveal differences between locutionary and illocutionary forces. For instance, Park (2022) claimed that '-e ju-' (to give) has recently been used

[1] The data can be obtained from https://corpus.korean.go.kr/request/reausetMain.do?lang=en and its processing can be conducted with the help of Korpora repository https://ko-nlp.github.io/Korpora/en-docs/corpuslist/modu_messenger.html

[2] Available in https://github.com/lovit/petitions_dataset

	Messenger	Petition/Vlog
# Sentences	6,407	125
Accuracy	85.92	52.87
Macro F1	51.61	22.44

Table 2: Speech act classification evaluation on two test sets of different domain (trained on the messenger dataset).

among Korean language users as a predicate to describe the behavior of the speaker her/himself, dominantly in the context of vlogging and microblogging. Also, owing to conventional pro-drop in Korean, this kind of phenomenon would make it much more difficult for trained models to infer the speech act just given a single utterance. Also, petition titles usually aim to appeal to the readers by using eye-catching phrases that include sarcasm (a representative figurative language where the user intention may differ from the locutionary force) or rhetorical questions, which also contribute to the classification difficulty.

Song (2023) took into account these kinds of language changes in contemporary Korean and addressed new criteria of Korean speech act categorization (Table 1). Speech acts were divided into five major categories following Austin (1962) and Searle (1976): *statement* that corresponds with declaratives, *suggestion* with directives, *exclamation* with exclamatives, *question* with interrogatives, and *greeting* with conventional expressions, with additional subcategories like *sarcasm/humor* and *rhetorical questions* added. We adopt these criteria for the annotation of datasets collected above; that is, we annotate the Messenger Corpus that consists of contemporary Korean colloquial utterances, use it for the model training, and check the model performance using all three types of sentences.

The annotation was conducted by computational linguists who have experience of Korean speech act annotation[3]. Especially for the test set with challenging examples (48 for vlogging expressions and 77 for petition titles), three Korean computational linguists participated in the annotation and obtained the Kappa of 0.715 (Fleiss, 1971)[4].

2.3 Experiment

For the automatic classification of speech acts, we adopted the bert-base-multilingual-cased model (Devlin et al., 2018) that utilized Wikipedia data for pre-training.

The model classification of speech acts underwent a fine-tuning process, a learning method conventionally used for PLM downstream tasks. The training set consists of randomly selected 25,611 instances (80%) of Messenger Corpus, while the test set incorporates 6,407 instances (20%) of it (batch size 32 with AdamW (Loshchilov and Hutter, 2017) optimizer). Accuracy and Macro F1 scores were used as evaluation metrics.

The classification accuracy for the messenger corpus was 85.92, with the F1 score 52.87 (Table 2). To verify whether the trained model adapts to comparably unseen expressions, a test conducted using the public petition titles and vlogging evaluation set (125 instances). We obtained the accuracy of 51.61 and F1 score 22.44, which implies that the model performance significantly differs from the validation with homogeneous dataset. It displays the discrepancy that comes from the domain difference of both types of sets.

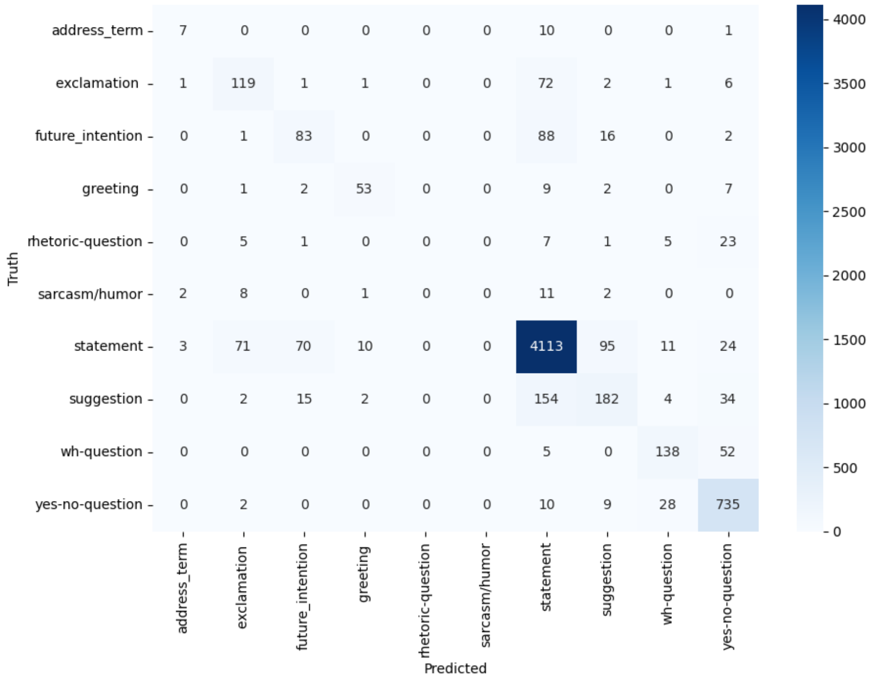
3 Analysis

3.1 Visualization and Error Analysis

To analyze the classification results, error rates among speech act categories were visualized through a heatmap generated via a confusion matrix (Figure 1), for the evaluation with Messenger Corpus (homogeneous to the training corpus). It was notably observed that the misclassification of *statements* as *suggestions* was prevalent, reaching 95 instances, and thus representing the most frequent misprediction. Furthermore, the error of classifying *future intentions* as *statements* was also significant, amounting to 88 instances. Overall, due to the high frequency of statements, the absolute frequency of misprediction involved *statements* being confused with *suggestions* or *future intentions* being misclassified as *statements*. Conversely, while not a high-frequency speech act, *rhetorical questions* demonstrated their trickiness, as the model did not accurately identify any instances, instead incorrectly categorizing them as yes-no questions in 23 instances. This exhibited a relatively high error rate in comparison to cases of accurate identification.

[3] The data and agreement/label are available only privately upon application due to the policy of the original providers.

[4] Available in https://github.com/songys/DAKoSA-Domain_Adaptation_in_Korean_Speech_Act



Truth \ Predicted	address_term	exclamation	future_intention	greeting	rhetoric-question	sarcasm/humor	statement	suggestion	wh-question	yes-no-question
address_term	7	0	0	0	0	0	10	0	0	1
exclamation	1	119	1	1	0	0	72	2	1	6
future_intention	0	1	83	0	0	0	88	16	0	2
greeting	0	1	2	53	0	0	9	2	0	7
rhetoric-question	0	5	1	0	0	0	7	1	5	23
sarcasm/humor	2	8	0	1	0	0	11	2	0	0
statement	3	71	70	10	0	0	4113	95	11	24
suggestion	0	2	15	2	0	0	154	182	4	34
wh-question	0	0	0	0	0	0	5	0	138	52
yes-no-question	0	2	0	0	0	0	10	9	28	735

Figure 1: Confusion matrix on the Messenger Corpus.

(2) Speaker 1: 이게 진정한 미식의 길이지 (*This is how a true foodie does it.*)

Speaker 2: ㅋㅋㅋ잘했다 폭식의 길 아닐까 (*Well done! Sounds like a road to gluttony to me.*)

Speaker 1: 조용히 해 줄래? ㅋㅋ (*Could you be quiet? lol*)

In example (2), '조용히 좀 해 줄래? ㅋㅋ (*Could you be quiet? lol*)' was interpreted as a *rhetorical question* by human annotators, but the model classified it as a *yes-no question*. In cases like the aforementioned example, humans might interpret the utterance variously as a *rhetorical question*, a *yes-no question*, or even an imperative, depending on the context. Such errors are presumed to stem from training the model at the sentence level without contextual information. Conversely, in the following example, both humans and the model successfully classified the utterance as a *rhetorical question*.

(3) Speaker 1: 와 가식쟁이다ㅋㅋㅋ (*Wow, what a hypocrite.*)

Speaker 2: 어쩌라고 죽을래? (*What are you gonna do about it, wanna die?*)

In the instance of example (3), responding with '죽을래' ('*wanna die?*') to the term '가식쟁이' ('*hypocrite*') poses a challenge to classify as either a *yes-no question* or an imperative. Thus, in clear contexts like this, both human annotators and the model aptly classified it as a *rhetorical question*, in contrast to situations where context is not provided and where the error rate appeared to be high due to interpretative challenges.

3.2 Further Analysis on RQs

A notable observation from the confusion matrix is that, in the case of *rhetorical questions*, out of 42 questions, 23 were annotated as *yes-no questions*, and 5 as a *wh-question*. It becomes evident that instances like *rhetorical questions*, where the overt sentence form and the underlying semantics differ, present heightened difficulties in classification.

Here, we discuss the case with examples from petition titles in which the model mispredicted a *rhetorical question* as a *yes-no question* (Table 3). Questions concern societally controversial topics in Korea, such as women's military service (which is not mandatory de facto), compensation issues for injuries during the service, and questions on murder and fundamental human rights issues. In these examples, humans annotated a question like "*Is it reasonable not to go to the army just because someone is female?*" not as a question necessitating a binary 'yes' or 'no' answer but as a *rhetorical question*, interpreting it as an emphatic expression. However, the model, probably not having been previously exposed to such types of questions (even in the Messenger Corpus where the sentences are

Example	Human annotation	Model prediction
여성이라는 이유만으로 군대를 안가는게 정상적인가요? (*Is it reasonable not to go to the army just because someone is female?*)	*rhetorical-question*	*yes-no-question*
군대에서 다쳤으면 국가가 보상해야 되지 않나요? (*Isn't it a duty of the nation to compensate for the injury in the army?*)	*rhetorical-question*	*yes-no-question*
부산 여중생 사건 이런 일 정말 반복 안될 수 없을까요? (*Couldn't we stop such a tragedy, like Busan middle schoolgirl incident?*)	*rhetorical-question*	*yes-no-question*
살인을 해야 살인자입니까? (*Do we only call someone a murderer only if he or she commits murder?*)	*rhetorical-question*	*yes-no-question*

Table 3: Petition examples where the model prediction differs from the human annotation.

daily conversation), categorized it as a *yes-no question*. One consideration that needs to be taken into account in a speech act analysis system is that a meticulous analysis of the domain of usage should precede before the inference.

The following example of vlogging text also represents a similar case.

(4) (김치찌개를 끓이는 영상)... 냄비에 채소 먼저 깔아 주고 김치를 반 포기 정도 ①**넣어 줍니다**. ...돼지고기 넣고 푹 ②**끓여 줄게요**. 고기는 목살이에요. (고기가 어느 정도 익은 후에) 먹기 좋은 크기로 ③**잘라 줍니다**. (각종 양념을 넣는다는 설명) 잘 섞어서 오래 ④**끓여 줄게요**....
*(In a video of cooking kimchi stew)... First, put the vegetables in the pot and then ①**add about half a head of kimchi**. ... Add pork and ②**simmer thoroughly**. The meat is pork neck. (After the meat has been cooked to some extent) ③**Cut it into** bite-sized pieces. (Explaining that various seasonings are added) Mix it well and ④**boil for a long time**. ...* (Park, 2022)

Example (4) above highlights a section from a vlog video wherein the speaker, a vlogger, is describing the ongoing process of a cooking activity s/he is engaged in. Notably, the speaker uses the '-어 주-(-e ju-)' expression, as in '넣어 줍니다' (add something) and '잘라 줍니다' (cut something), wherein the agent and the beneficiary of the action reside in the same clause.

So far, in the Korean language, these expressions have not been used by language users to describe the behavior of the speaker her/himself. In this regard, in the experiment using vlogging script, the model predicted 5 out of 6 items as *suggestions* in instances for the pro-drop cases (frequent in Korean spoken language), and predicted as *statements* when the subject was explicitly stated. In other words, the intention of these types of utterances can be determined upon the viewpoint and timestamp of the analysis; the vlogger would have said the utterance with an intention of describing his/her behavior, but the audience of the vlog would interpret it as a suggestion of cooking sequences. This implies that, particularly in pro-drop languages like Korean, a correct understanding of utterance intent may be possible if and only if an accurate and contextual speech act annotation is performed, which reflects the importance of not only domain but also cultural and time-variant characteristics.

4 Conclusion

In this study, speech acts were annotated on the NIKL Messenger Corpus, the titles of public petitions, and vlogging scripts, focusing on the analysis of error items in sentences with discrepancy between locutionary and illocutionary force. Additionally, it turned out that stylistic features used in a specific circumstances also influence the decision of speech acts. Considering different contexts, such as messenger conversations that belong to private dialogue and public petitions or vlogging script that have the nature of having the audience, it is deemed necessary to build and analyze balanced corpora across various domains concerning whether the discourse is public or not and having multiple or anonymous addressee.

Acknowledgments

We thank anonymous reviewers for constructive comments. Authors are also grateful for Jisu Shim for helping the annotation process.

References

J. L. Austin. 1962. *How to do Things with Words*. Oxford University Press, New York. Reprinted 1975.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum.

2010. Towards an ISO standard for dialogue act annotation. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Languages Resources Association (ELRA).

Won Ik Cho and Nam Soo Kim. 2022. Text implicates prosodic ambiguity: A corpus for intention identification of the korean spoken language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(1):1–20.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

NIKL. 2020. Nikl corpora 2020 (v.1.0).

Jinho Park. 2019. Yes-no-questions and rhetorical questions. *Hangughagbo*, pages 1–24.

Mi-eun Park. 2022. About the '-e ju-' construction of that i use to myself –focused on vlog register–. *Korean Semantics*, 78:89–117.

John R Searle. 1976. A classification of illocutionary acts1. *Language in society*, 5(1):1–23.

Sanghoun Song. 2010. Pragmatic usage of wh-elements in korean. *Language Information*, 11:91–113.

Youngsook Song. 2023. *Enhancing AI's Commonsense Reasoning in Conversations through Natural Language Generation*. Ph.D. thesis, University of Kyunghee.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Readability and Complexity: Diachronic Evolution of Literary Language Across 9000 Novels

Yuri Bizzoni
Center for Humanities Computing
Aarhus University, Denmark
yuri.bizzoni@cc.au.dk

Pascale Feldkamp
Center for Humanities Computing
Aarhus University, Denmark
pascale.moreira@cc.au.dk

Ida Marie Lassen
Center for Humanities Computing
Aarhus University, Denmark
idamarie@cas.au.dk

Mads Rosendahl Thomsen
Comparative Literature
School of Communication and Culture
Aarhus University, Denmark
madsrt@cc.au.dk

Kristoffer Nielbo
Center for Humanities Computing
Aarhus University, Denmark
kln@cas.au.dk

Abstract

Using a large corpus of English language novels from 1880 to 2000, we compare several textual features associated with literary quality, seeking to examine developments in literary language and narrative complexity through time. We show that while we find a correlation between the features, readability metrics are the only ones that exhibit a steady evolution, indicating that novels become easier to read through the 20th century but not simpler. We discuss the possibility of cultural selection as a factor and compare our findings with a subset of canonical works.

1 Introduction

Several textual features have been associated with "good style" or narrative in the stylometric and quantitative literature studies. A recent surge of quantitative studies has used large corpora to investigate whether intra-textual features correlate with perceived literary quality. Average sentence length (Ganjigunte Ashok et al., 2013), type-token ratio (Crosbie et al., 2013), the distribution of parts of speech (van Cranenburgh and Bod, 2017) and level of redundancy (Algee-Hewitt et al., 2018) have been shown to explain literary success partially.

Also measures of readability are often connected to literary success: it is a widespread conception of both readers and publishers that bestsellers are easier to read (Martin, 1996), and readability has recently gained traction in creative writing and publishing, such as in text-editing tools like the Hemingway[1] or Marlowe editors. [2] These applications evaluate texts with simple readability measures and tend to encourage the production of texts that are easier to read, assuming that more readable texts are better.

With the evolution of quantitative methodologies, more sophisticated models of texts have also been explored as possible markers of literary quality: the shape and dynamics of novels' sentiment arcs as a way to approximate their narrative development or the complex way parts of speech alternate throughout a text, influencing readers' experience of the story above or below conscious perception (Bizzoni et al., 2021, 2022b; Mohseni et al., 2022).

Literary evolution may show a progressive convergence towards preferred forms of and styles in narrative, perceived as effective and maintained/further evolved through community feedback (Crocker et al., 2016; Degaetano-Ortlieb and Teich, 2022). Already in the 19th century Sherman (1893) observed an evolution of the language of fiction and suggested a positive selection for simple language in literary language. This idea recurs in a theory where the rise of a mass readership is thought to have prompted the language

[1] https://hemingwayapp.com/help.html

[2] https://authors.ai/marlowe/

of Western fiction to become simpler through the 19th and 20th centuries, as it caters to the progressively lower overall literacy and less spare time of the readership (Klancher, 1983; Kimball, 2017; Westin, 2002).

In this study, we first extract multiple textual and stylometric measures connected to perceived literary quality or success from a large collection of English novels, ranging from the most surface-level readability indices to models considering the dynamics of their sentiment arcs. We examine whether any systematic, diachronic trend of these measures can be observed in the period covered by our corpus (1880-2000), as has been noted in theories of slow but continuous change in literary language or narrative style into the 20th century (Underwood, 2019; Underwood and Sellers, 2016; Moretti, 2000). Secondly, considering readability as a measure linked to literary success or quality, we test the correlation between readability and other stylistic measures, as well as two measures with a higher level of abstraction that have previously been used to estimate narrative complexity in relation to reader appreciation: fractality and entropy. These measures are based on the sentiment arcs of novels, which are the sentiment scores (often extracted through dictionaries or machine learning) over the course of a whole novel. Certain shapes or sentiment arc dynamics have been connected to reader appreciation, considering both simple and more complex narratives (Bizzoni et al., 2022a), and Bizzoni et al. (2023) have shown that sentiment features, such as measures of sentiment arc progression, have an effect even compared to the predominantly stylistic features usually employed for this type of task (Koolen et al., 2020; Maharjan et al., 2017). These more complex measures that take into account the sentiment-arc of novels are interesting insofar as they are not direct measures of style, and insofar as they have proven effective in approximating literary quality for different types of quality-standards that may reflect tastes of different reader communities: distinguishing higher-rating works on large user-platforms such as GoodReads (Bizzoni et al., 2021) and telling works of Nobel laureates from those of contemporary authors (Bizzoni et al., 2022b). Finally, we estimate the same measures and diachronic trend for a subsection of the corpus defined through a combination of different "quality resources" chosen to reflect canonicity: authors most often appearing in English Literature syllabi, major literary anthologies and titles defined as "classics" on the large user-platform GoodReads. We show that while there is a correlation between surface-level and arc-based metrics, their change through time is significantly different, with readability metrics being time-dependent, and more sophisticated measures time-independent.

2 Related works

2.1 Readability and text complexity

The connection between text readability and quality has often been implied for non-fiction. Early studies of readability attest to the educational and social concerns in developing measures of readability to improve expository or didactic texts (Chall, 1947). Yet, the role of readability in the quality of *literary* texts is a more complex question, where "opacity" has also been considered a positive trait (Glissant, 1997; White et al., 1981; Moore, 1964).

Few studies have examined readability measures for predicting literary quality or success. Studying a small corpus of bestsellers and more canonical literary works, Martin (1996) found no significant difference in readability using a modified Flesch Reading Ease. In contrast, Garthwaite (2014) found differences in readability between bestsellers and commercially endorsed book-list titles, where endorsed lists of books were more difficult to read. Relying on multiple measures of readability and one measure of literary quality (i.e., GoodReads' average ratings), Maharjan et al. (2017) found that readability was not effective for estimating popularity when comparing it to, for example, character n-grams. Similarly Koolen et al. (2020) preferred other features over readability measures when estimating perceived literary quality. Still, many studies of literary success, popularity, or perceived literary quality have sought to approximate text complexity and have studied textual properties upon which formulae of readability are directly or indirectly based, such as sentence length, vocabulary richness, and text compressibility (Brottrager et al., 2022; van Cranenburgh and Bod, 2017; Crosbie et al., 2013).

2.2 Sentiment arcs

More complex measures based on the linear development of novels – their sentiment arcs – have been used to approximate literary quality for different types of reader standards, and estimate narrative rather than style. The sentiment or emotion-based development of communication is often seen as highly relevant, especially in "artistic" narrative (Drobot, 2013), as it is linked to the central and special tendency of literary texts to evoke, and not only describe experiences and inner states. As Hu et al. (2021) argues, readers engage with the evolution of a story at the emotional level by evocations or "engagement prompts". A sentiment arc is thus a model of the "engagement prompts" in the text, which sentiment analysis models as a primary tool approximate at the word (Mohammad, 2018), sentence (Mäntylä et al., 2018) or paragraph (Li et al., 2019) level (Cambria et al., 2017; Kim and Klinger, 2018; Brooke et al., 2015; Jockers, 2017; Alm, 2008; Jain et al., 2017). Sentiment analysis usually derives its models from human-based resources such as annotated lexica (Mohammad and Turney, 2013) or lists of words induced from labelled documents (Islam et al., 2020). Several studies have also attempted to complement the simplicity of sentiment analysis with systems for textual emotion recognition (Alm et al., 2005), or by developing more complex sentimental tools (Xu et al., 2020). In general, researchers have looked at sentiment arcs in terms of their overall shape (Reagan et al., 2016), but recent works have tried more complex mathematical models to define the arcs' overall level of inner coherence and predictability (Gao et al., 2016). In this study we recur to this last form of series modeling, examining the dynamics of arcs.

2.3 Quality measures

Defining literary quality as one unified standard and formalizing it for quantitative studies is a particularly complex and elusive problem. Studies that seek to predict perceived literary quality from textual features often rely on the provisional proxy of one single gold standard, such as book ratings from large user platforms such as GoodReads (Maharjan et al., 2018; Bizzoni et al., 2021) - usually with the task of predicting high-rated works - or personally as well as institutionally compiled canons (Mohseni et al., 2022), sales-numbers (Wang et al., 2019), or occasionally selections from prestigious awards such as the Nobel prize (Bizzoni et al., 2022b). However, it has been shown that reader preferences are complex and reflect multiple perceptions or standards of quality (Koolen et al., 2020), that are not necessarily based on the same criteria or prompted by the same textual features. For the present work, we use different standards of literary prestige that reflect a particularly "canonical" literary quality as a subset of our corpus to test against the wider set of titles.

3 Methods

3.1 Data

This present study uses the *Chicago corpus*, a collection of over 9,000 novels written or translated into English, spanning from 1880 to 2000. The titles were selected based on the number of libraries holding a copy of the novel (see Table 1).

The collection is rare in terms of its diversity - it represents well-known genres and popular fiction as well as important works from the entire period, including seminal modernist and postmodernist texts as well as Nobel Prize winners and recipients of prestigious literary awards. As such, the Chicago corpus contains a sizeable subsection of prestigious or "canonic" literature.

To estimate the amount of "canonic" literature in the corpus we mark all titles by authors that appear in selected institutional or user-compiled resources indicating literary prestige: in the English and American Norton Anthology (Shesgreen, 2009), two GoodReads user-generated lists, "the GoodReads classics" and the GoodReads "best books of the 20th century" (Walsh and Antoniak, 2021), and among the top thousand most assigned titles in English Literature course syllabi.[3] The amount of these “canonic" titles through time is shown in Fig. 1.

It should be noted that the Chicago corpus contains only works that were either produced or translated into English, exhibiting a clear cultural and geographic bias with a strong over-representation of Anglophone authors. This should also be considered in light of the fact that the readability metrics we use are particularly effective and were developed for the English language.

	Titles	**Authors**
Number	9089	3150
Avg. holdings	535.73	495.1

Table 1: Overall number of titles and authors in the corpus (first line) and average number of library holdings per title and per author (second line).

[3] Based on syllabi collected by the Open-syllabus project: https://opensyllabus.org

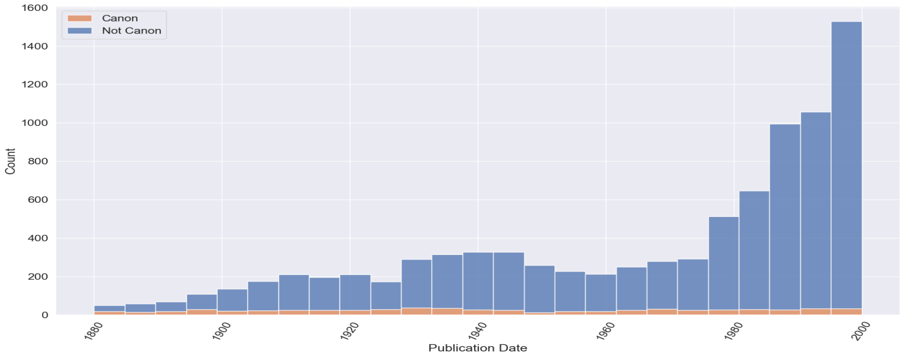


Figure 1: Overall quantity of titles per decade in the corpus, with the number of "canonical" books in orange.

Resource	N. titles
University Syllabi	478
Norton Anthology	402
GoodReads classics	62
GoodReads 20th century	44
Total unique titles	641

Table 2: Number of titles per canonicity resource in the Chicago corpus.

3.2 Measures of readability

While what is "readable" is problematic to define, and clearly varies depending on the reader, the context and the genre (Berlatsky, 2015; Flesch, 1948), readability scores may act as proxy measures for people's reading experience and enable comparison between texts.[4]

To avoid relying on one single interpretation of the readability concept, we compare five different measures of textual readability, chosen for their popularity and interpretability.[5] Since the 1920s, and particularly after the success of Flesch and Dale-Chall formulas in the 1950s, combinations of sentence length, word lengths, and/or number of syllables have been used as proxies for linguistic complexity to gauge the difficulty of a text (Dale and Chall, 1948). In 1980, there were more than 200 distinct readability formulae (Dubay, 2004), and new ones are continually being developed as older ones are refined. Despite their relative simplicity, the measures from what Dubay (2004) refers to as the "classic readability" studies remain the most popular ones and useful in determining text difficulty (Stajner et al., 2012).

The selected readability measures are the following:

- The *Flesch Reading Ease* is a measure of readability based on the average sentence length (ASL), and the average number of syllables per word (ASW). It is calculated as follows:

$$\begin{aligned}\mathrm{RE} = 206.835 - (1.015 \times \mathrm{ASL}) \\ - (84.6 \times \mathrm{ASW})\end{aligned}$$

- The *Flesch-Kincaid Grade Level* is a revised version of the Flesch Reading Ease score. Like the former, it is based on the average sentence length (ASL), and the number of syllables per word (ASW). It is calculated as follows:

$$\mathrm{GL} = (0.4 \times \mathrm{ASL}) + (12 \times \mathrm{ASW}) - 15$$

- The *SMOG Readability Formula* is a readability score introduced by McLaughlin (McLaughlin, 1969). It measures readability based on the average sentence length and number of words with more than 3 syllables (number of polysyllables), applying the formula:

$$\begin{aligned}\mathrm{SMOG} = 1.043 \times \sqrt{polysyllablecount} \\ \times \frac{30}{\mathrm{n_{st}} + 3.1291}\end{aligned}$$

- The *Automated Readability Index* is a readability score based on the average sentence

[4] We use the term "readability" here, since this is properly what readability indices, developed in linguistics, intend to measure. Other terms, like text "simplicity" may be related but are more broadly defined and often measured with a combination of both stylistic and more content-based features (Popović et al., 2022), while "readbaility" is predominantly stylistic.

[5] All readability scores were extracted using the textstat package: https://pypi.org/project/textstat/

length and number of characters per words (word length). It is calculated as follows:

$$\text{ARI} = 4.71\frac{\text{characters}}{\text{words}} + 0.5\frac{\text{words}}{\text{sentences}} - 21.43$$

- The *New Dale–Chall Readability Formula* is a 1995 revision of the Dale-Chall readability score (Chall and Dale, 1995). It is based on the average sentence length (ASL) and the percentage of "difficult words" (PDW) defined as words which do not appear on a list of words which 80 percent of fourth-graders would know (Dale and Chall, 1948).[6] It is calculated as follows:

$$\begin{gathered}\text{DC} = 0.1579 \times \text{PDW} + 0.0496 \times \text{ASL} \\ \text{If PDW} > 5\% : \text{Adjusted Score} = \\ \text{Raw Score} + 3.6365\end{gathered}$$

We complement these standard readability metrics with three other metrics often used to assess stylistic complexity of texts:

- **Sentence length**. Character-based sentence length is also often integrated into readability measures.

- **Type-token ratio**. A standard index of lexical richness, not used in readability metrics but normally considered indicative of a text's complexity and inner diversity (Torruella and Capsada, 2013).[7]

- **Compressibility** measures to what extent a text can be compressed through a standard compression algorithm. This measure becomes essentially a sign of redundancy or formulaicity: the more a text tends to repeat sequences *ad verbatim*, the more compressible it will be (Benedetto et al., 2002; van Cranenburgh and Bod, 2017).[8]

3.3 Sentiment arcs

To apply more complex, sentiment arc based metrics of the narratives, in this study, we extract sentiment arcs using the VADER model (Hutto and Gilbert, 2014) at the sentence level. Sentiment analysis of a literary text provides a simple and intuitive representation of a narrative's sentimental trajectory, and has been applied as a proxy for meaningful aspects of the reading experience (Drobot, 2013; Cambria et al., 2017; Kim and Klinger, 2018; Brooke et al., 2015; Jockers, 2017). The resulting representation is referred to as a sentiment arc, which a range of studies model to evaluate narratives in terms of genre (Kim et al., 2017), plot archetypes (Reagan et al., 2016), and lastly, reader preference (Bizzoni et al., 2022a). While dictionary-based sentiment analysis remains a popular choice, more recent, transfomer-based methods are more recently explored (Elkins, 2022).

Our choice of a dictionary-based approach was motivated by a desire for transparency and corpus independence. Among available sentiment analysis tools we selected VADER due to its widespread usage and comprehensive rule set. VADER generates a compound valence score for each sentence, ranging from negative (-1), through neutral (0), to positive (1). Figure 2 serves as a demonstration of the arc extraction process for the first ten sentences of Ernest Hemingway's seminal work *The Old Man and the Sea*. [9] To highlight the efficacy of the annotation on narrative texts, Figure 3 also shows the sentiment arc of *The Old Man and the Sea* with its corresponding narrative events, compared to human annotation of the book.[10]

[6] Contained in the Dale-Chall word-list: https://countwordsworth.com/download/DaleChallEasyWordList.txt

[7] We used a common method insensitive to text-length: the Mean Segmental Type-Token Ratio (MSTTR). MSTTR-100 represents the overall average of the local averages of 100-word segments of each text.

[8] We calculated the compression ratio (original bit-size/compressed bit-size) for the first 1500 sentences of each text using bzip2, a standard file-compressor.

[9] "He was an old man who fished alone in a skiff in the Gulf Stream and he had gone eighty-four days now without taking a fish. In the first forty days a boy had been with him. But after forty days without a fish the boy's parents had told him that the old man was now definitely and finally salao, which is the worst form of unlucky, and the boy had gone at their orders in another boat which caught three good fish the first week. It made the boy sad to see the old man come in each day with his skiff empty and he always went down to help him carry either the coiled lines or the gaff and harpoon and the sail that was furled around the mast. The sail was patched with flour sacks and, furled, it looked like the flag of permanent defeat. The old man was thin and gaunt with deep wrinkles in the back of his neck. The brown blotches of the benevolent skin cancer the sun brings from its reflection on the tropic sea were on his cheeks. The blotches ran well down the sides of his face and his hands had the deep-creased scars from handling heavy fish on the cords. But none of these scars were fresh. They were as old as erosions in a fishless desert. Everything about him was old except his eyes and they were the same color as the sea and were cheerful and undefeated."

[10] Human annotation of sentiment per sentence was performed by 2 annotators, asked to score individual sentences

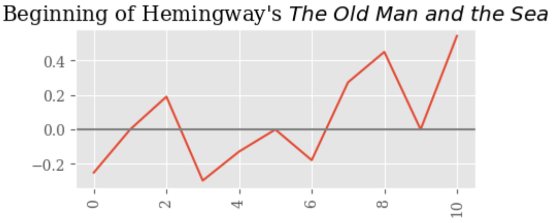


Figure 2: VADER-annotation of Hemingway's *The Old Man and the Sea*, the first 10 sentences. The arc captures the narrative fluctuations of the sentence sequence.

3.4 Sentiment arc-based metrics

To model the underlying complexity of the novels, we use two more complex measures already previously used in studies of literary quality: Hurst's exponent and Approximate Entropy of the novels' sentiment arcs (Bizzoni et al., 2021, 2022b; Mohseni et al., 2022).

Used to detect long-term memory in time series data, the Hurst exponent in our context measures the persistence of sentiment or the long-term memory of sentiment arcs. To estimate Hurst, we combine non-linear adaptive filtering with fractal analysis, specifically adaptive fractal analysis (Gao et al., 2011; Tung et al., 2011). Nonlinear adaptive filtering is used due to the inherent noisiness of story arcs. First, the signal is partitioned into segments (or windows) of length $w = 2n + 1$ points, where neighboring segments overlap by $n + 1$. Then, a polynomial of order D is fitted for each segment. The fitted polynomial for ith and $(i + 1)th$ is denoted as $y^{(i)}(l_1), y^{(i+1)}(l_2)$, where $l_1, l_2 = 1, 2, ..., 2n + 1$. We use the following weights for the overlap of two segments.

$$y^{(c)}(l_1) = w_1 y^{(i)}(l + n) + w_2 y^{(i)}(l), \quad l = 1, 2, \ldots, n + 1 \quad (1)$$

where $w_1 = (1 - \frac{l-1}{n}), w_2 = 1 - w_1$ can be written as $(1 - \frac{d_j}{n}), j = 1, 2$, where d_j denotes the distance between the point of overlapping segments and the center of $y^{(i)}, y^{(i+1)}$. Studies have demonstrated the usefulness of adaptive filtering applied to sentiment arcs, especially in the context of estimating dynamics of sentiment arcs (Hu et al., 2021; Bizzoni et al., 2022b).

After nonlinear adaptive filtering, we use the Hurst exponent to measure long-term memory. Assuming that stochastic process $X = X_t : t = 0, 1, 2, ...$, with stable covariance, mean μ and σ^2, the process' autocorrelation function for $r(k), k \geq 0$ is:

$$r(k) = \frac{E\left[X(t)X(t+k)\right]}{E\left[X(t)^2\right]} \sim k^{2H-2}, \text{as} \quad k \to \infty \quad (2)$$

where H is called the Hurst exponent (Mandelbrot, 1982).

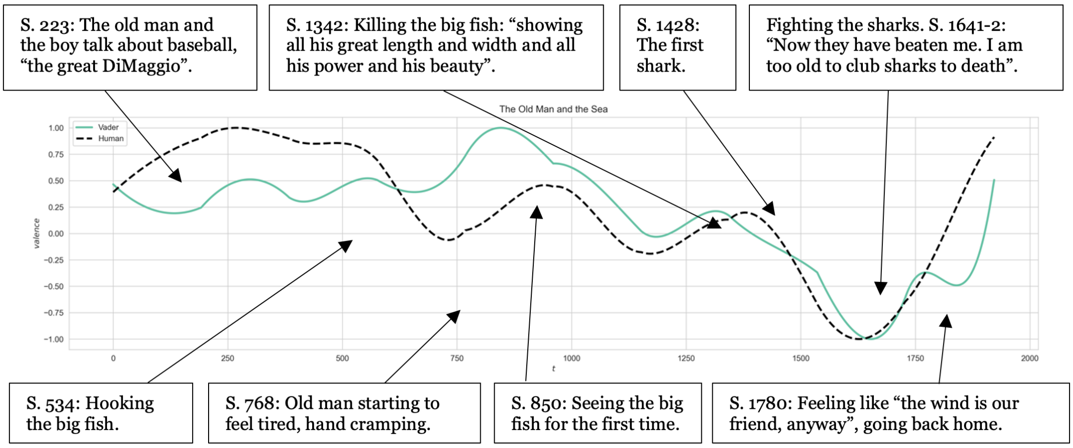
For $0.5 < H < 1$ the story arc is characterized as persistent such that increments are followed by increases and decreases by further decreases. For $H = 0.5$ the story arc only has short-range correlations; and when $H < 0.5$ the story arc is anti-persistent such that increments are followed by decreases and decreases by increments. For the specific application domain (i.e., narratives) persistent story arcs are characteristic of coherent narratives, where the emotional intensity evolves at longer time scales. Story arcs that only show short memory lack coherence and appear like a collection of short stories. Anti-persistent story arcs will appear bland and rigid narratives oscillating around an average emotional state (Hu et al., 2021).

Adaptive fractal analysis consists of the following steps: first, the original process is transformed to a random walk process through first-order integration $u(n) = \sum_{k=1}^{n}(x(k) - \overline{x}), n = 1, 2, 3, ..., N$, where $\overline{x}$ is the mean of $x(k)$. Second, we extract the global trend $(v(i), i = 1, 2, 3, ..., N)$ through the nonlinear adaptive filtering. The residuals $(u(i) - v(i))$ reflect the fluctuations around a global trend. We obtain the Hurst parameter by estimating the slope of the linear fit between the residuals' standard deviation $F^{(2)}(w)$ and w window size as follows:

$$F^{(2)}(w) = \left[\frac{1}{N}\sum_{i=1}^{N}(u(i) - v(i))^2\right]^{\frac{1}{2}} \sim w^H \quad (3)$$

Beyond Hurst exponent, we estimate the approximate entropy (ApEn) of sentiment arcs. ApEn is a measure of the predictability of time-series of data based on Shannon Entropy and introduced by S. Pincus as a measure of physiological system complexity (Pincus, 1991; Pincus et al., 1991). Given a time series X with N data points, ApEn is calculated as follows: a value for m, the length of the comparison segment, and a tolerance value r are chosen. The time series X is then divided

of the book on a 1-10 scale without paying attention to the narrative context.



S. 223: The old man and the boy talk about baseball, “the great DiMaggio”.

S. 1342: Killing the big fish: “showing all his great length and width and all his power and his beauty”.

S. 1428: The first shark.

Fighting the sharks. S. 1641-2: “Now they have beaten me. I am too old to club sharks to death”.

S. 534: Hooking the big fish.

S. 768: Old man starting to feel tired, hand cramping.

S. 850: Seeing the big fish for the first time.

S. 1780: Feeling like “the wind is our friend, anyway”, going back home.

Figure 3: The detrended (by adaptive filtering) and normalized sentiment arcs of *The Old Man and the Sea* based on VADER scores and human annotations, shown with main narrative events.

into overlapping segments of length m, such that X_i to X_{i+m-1} represents one segment, where $1 \leq i \leq N - m + 1$. For each segment X_i to X_{i+m-1}, the number of segments X_j to X_{j+m-1} (where $j \neq i$) that are within a distance of r from X_i to X_{i+m-1} is calculated, where r is a real number that specifies a filtering level, essentially defining what constitutes a match. We will call the number of matches $C(i)$. Finally, the probability of observing $C(i)$ matches for a given segment X_i to X_{i+m-1} as can be computed as:

$$p(i) = \frac{C(i)}{N - m + 1} \tag{4}$$

The previous steps are be repeated for increasing values of m and the probabilities are averaged over all segments to obtain the final value:

$$ApEn(m, r) = -\frac{1}{N - m + 1} \sum [\ln (p(i))] \tag{5}$$

The ApEn value for a given time series is determined by the minimum value of $ApEn(m, r)$ for a range of m and r values.

The ApEn value represents the level of randomness or predictability in the time series, with higher values indicating greater randomness and lower values indicating more predictability. ApEn has been used to study the complexity of various types of time series data, i.a., heart-rate (Fleisher et al., 1993), financial (Delgado-Bonal and Marshak, 2019), and narratives (Mohseni et al., 2022). Applied to sentiment arcs, ApEn searches for recurrent patterns in the arc and estimates the (log) likelihood that adjoining sequences of sentences, two in this study, will differ, that is, whether the pattern is predictable. Smaller values of ApEn indicate more recurring patterns and thus higher predictability, while higher values indicate fewer recurring patterns and lower predictability.[11]

4 Results

As we show in Figure 4 and Table 3, the main result of our analysis consists of two series of trends:

1. All measures of readability clearly correlate with the passage of time and point in the same direction: to an increased readability of novels. Sentence length follows this trend, indicating that sentences become on average shorter through the 20th century.

2. All other measures we took into consideration, including the "linear" measures of sentiment arcs, do not change meaningfully through time.

The clear trend of all readability measures indicates an overall simplification of the literary prose, beyond the characteristics of the authors’ individual styles. Interestingly, the trend can be observed for the corpus at large as well as for the "high prestige" subsection of titles we outline in Table 2. Looking into the relation between the readability measures and Hurst as well as Approximate Entropy (Table 4) we find that they correlate with readability in the sense that more difficult books tend to have a higher Hurst exponent and higher Approximate Entropy. So overall, in our corpus, simpler books have

[11] We used the Neurokit-package to measure ApEn of arcs: https://neuropsychology.github.io/NeuroKit/

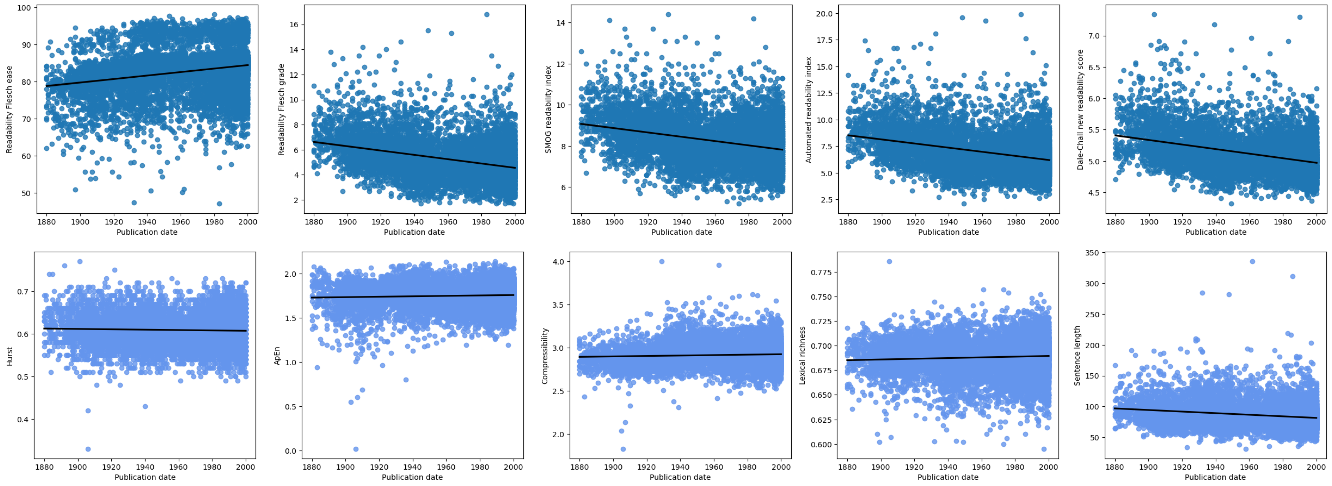


Figure 4: Distribution of readability measures (upper) and other measures (lower row) through time. Note that Flesch Reading Ease shows a score where a lower number means lower readability so that it is inverted with respect to the other readability measures.

less complex arcs. However, we do not see a tendency towards simpler arcs through time: if books become easier to read from 1880 to 2000, they do not become simpler in terms of their sentiment-arc dynamics. The overall level of complexity of the novels' sentiment arcs remains remarkably stable through the corpus - and the titles of our "canon selection" even show a slight tendency towards higher complexity through time. While this points to the fact that readability and arc complexity are only partially correlated (other factors might correlate even more strongly with one or the other), it shows that with time writers might have increasingly favored a kind of prose that strives to keep a non-obvious balance between simplicity of style and complexity of the sentiment arc.

The lack of lasting diachronic changes in the other two stylistic measures considered, type-token ratio and textual compressibility, seems to confirm this picture: if novels become easier in terms of basic readability metrics, they do not lose complexity at many other levels, not becoming overall more repetitive nor lexically poorer. In other words, it might be that there has been a large, overall tendency to favor texts that manage to simplify the most surface level aspect of style, without compromising their linguistic diversity nor their narrative arcs' complexity.

	Spearman	**Pearson**
Hurst	-0.015 (*-0.1*)	0.036
ApEn	0.032 (*0.1*)	0.05
Lexical richness	0.081 (*0.0*)	0.062
Compressibility	0.042 (*-0.1*)	0.006
Sentence length	-0.185 (*-0.1*)	-0.201
Flesch Ease	0.249 (*0.2*)	0.246
Flesch Grade	-0.316 (*-0.3*)	-0.362
SMOG	-0.287 (*-0.2*)	-0.323
ARI	-0.296 (*-0.3*)	-0.352
Dale Chall	-0.341 (*-0.2*)	-0.383

Table 3: Spearman and Pearson correlations between textual measures and the novels' publication date. For reference, the Spearman correlations of textual meaures and the novel's publication date for *canonic works only* are added in parentheses. All non-null correlations (r>0.1) have p-values < 0.0005.

	Hurst	**ApEn**
Hurst	1.000	0.366
ApEn	0.366	1.000
Flesch Ease	-0.162	-0.404
Flesch Grade	0.172	0.431
SMOG	0.153	0.412
ARI	0.172	0.428
Dale Chall	-0.043	0.104

Table 4: Spearman correlations between linear metrics and readability measures (all statistically significant).

5 Discussion

The different trends we have shown between surface-level readability measures, other metrics of style, and arc complexity through time seem to point towards a large-scale evolution of literary language towards prose that favors increased readability without compromising the novels' linguistic or narrative versatility. As we have seen in Table 4, arc measures (Hurst exponent and Approximate Entropy) and readability are indeed correlated in the corpus, but the development of readability measures shows a tendency to progressively simplify the prose of all novels, including those with complex arcs (Table 3).

Regarding the trends towards readability alone, it is reasonable to exclude that they are the effect of an overall change of the English language. Similar tendencies towards simplification have been found in narrative (Sherman, 1893; Liddle, 2019) but is not as obvious in other domains (Säily et al., 2017). Moreover, scientific and journalistic prose has even shown an opposed trend, with texts becoming more difficult to read (Plavén-Sigray et al., 2017; Danielson et al., 1992).

If this trend is not an effect of language change, its presence in literature can give way to intriguing hypotheses. The emergence of what scholars have called mass readership (Klancher, 1983) and a widening of the alphabetized population might have pushed the success of easier books,[12] while the increasingly pressing market logic applied to the editorial world might have helped shaping literary style into simpler and simpler forms, easier to consume in a shorter time (Winter and O'Neill, 2022). It is also possible that in the last century, difficulty of reading has shifted from a virtue to a vice in the view of the English writing world, with novelists and publishers alike slowly favoring more direct or transparent prose.

A central question to ask is whether we are seeing an actual transformation of literary prose, or whether we are witnessing an effect of cultural selection. In theory, there might have been no evolution at all through the 120 years in question, but less appreciated exemplars might have been progressively lost or overlooked. If texts are undergoing a constant process of selection, it is possible that the "books worth keeping" maintain more complex stylistic features, while the larger number of more easily readable novels is progressively forgotten, leading to the illusion of a historical change through survivor bias. A look at the absolute number and percentage of "canonic" books in our corpus, as defined by various indicators of prestige (see Section 3.1), seems to point to this competing view: while the number of texts in the corpus increases with each decade, the percentage of canonic titles decreases drastically through time.

However, when looking at the canonic subset alone, we see changes similar to those that we observe in the whole corpus: what we have defined as canonical literature has also become more readable from 1880 to 2000 (Table 3). Moreover, the canonic subset seems to tend even more toward a disentanglement of surface readability and arc complexity, with the latter showing even a slight increase in complexity – Hurst and ApEn – through time.

6 Conclusion and Future Works

In our analysis of a curated corpus of 9000 English-language novels published between 1880 and 2000, we employed specific readability metrics – i.a., the Flesch-Kincaid Readability Score – as well as complexity indices like the Hurst exponent of sentiment arcs and classic stylometrics, i.a., type-token ratio and compressibility. Our data indicates some clear trends: most readability scores have increased through time, displaying Spearman correlations of up to 0.34 with the publication year, signifying a gradual simplification of the overall narrative language. In contrast, richness and complexity metrics remained relatively unchanged over the same period. These divergent trends might suggest that authors are increasingly focusing on making their works more accessible while maintaining a consistent level of narrative complexity and lexical diversity, which we might interpret as a literary strategy to engage a broader audience without sacrificing depth or complexity. It's worth noting that our study does not account for genre-specific trends and is based on available works, thus introducing potential selection bias.

Future research could expand upon these findings by exploring how these trends vary across different genres and cultures. Naturally, exploring this further would require properly discussing and deploying a system of judgments for literary quality, an undertaking beyond the scope of this work. In the future, we would like to both conduct qualitative analyses to assess these results on individual work level and repeat the experiment on different and possibly larger literary data sets. We also plan to collect more textual features, such as model perplexity, as well as develop more sophisticated models for the Sentiment Analysis that underlies measures of arc dynamics (Hurst, ApEn), such as using LMs, and examine the change through time of these features with more sophisticated mathematical models.

[12] The US National Reader Survey in 1993 found that 48 percent of the adult population have difficulties reading above 5th-grade level texts (Kirsch et al., 1993)

Limitations

The Chicago Corpus serves as a valuable resource for our study, as it encompasses an expansive and representative sample of widely read Anglophone literature over a century, allowing for a robust analysis. Still, it is worth noting that the corpus has a geographical bias: most authors are of US origin and few are non-Anglophone. This bias inevitably situates the entire analysis within the context of a well-defined “Anglocentric" literary field. Moreover – perhaps also due to an inherent skew in this literary field – 36% of authors are women.

While these imbalances do not inherently undermine our experiments, it is crucial to bear them in mind when interpreting the results, and we advise against extrapolating our findings to the context of a wider or global literary field. Moreover, when estimating the canonicity of works in the corpus we have relied on external lists that are, to an extent, characterised by similar biases, for example the *Norton Anthology*, which is similar both in terms of most predominantly selecting among Anglophone authors and in terms of its gender biases (Pope, 2019). We trust that any interpretation of our findings will have these limitations in mind.

References

Mark Algee-Hewitt, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, and Hannah Walser. 2018. Canon/archive : large-scale dynamics in the literary field.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 579–586.

Ebba Cecilia Ovesdotter Alm. 2008. *Affect in text and speech*. University of Illinois at Urbana-Champaign.

Dario Benedetto, Emanuele Caglioti, and Vittorio Loreto. 2002. Language Trees and Zipping. *Physical Review Letters*, 88(4):1–5.

Noah Berlatsky. 2015. Readability is a myth. Section: Culture.

Yuri Bizzoni, Pascale Moreira, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2023. Sentimental matters - predicting literary quality by sentiment analysis and stylometric features. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 11–18, Toronto, Canada. Association for Computational Linguistics.

Yuri Bizzoni, Telma Peura, Kristoffer Nielbo, and Mads Thomsen. 2022a. Fractal sentiments and fairy tales-fractal scaling of narrative arcs as predictor of the perceived quality of andersen's fairy tales. *Journal of Data Mining & Digital Humanities*.

Yuri Bizzoni, Telma Peura, Kristoffer Nielbo, and Mads Thomsen. 2022b. Fractality of sentiment arcs for literary quality assessment: The case of nobel laureates. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 31–41, Taipei, Taiwan. Association for Computational Linguistics.

Yuri Bizzoni, Telma Peura, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2021. Sentiment dynamics of success: Fractal scaling of story arcs predicts reader preferences. In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 1–6, NIT Silchar, India. NLP Association of India (NLPAI).

Julian Brooke, Adam Hammond, and Graeme Hirst. 2015. Gutentag: an nlp-driven tool for digital humanities research in the project gutenberg corpus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 42–47.

Judith Brottrager, Annina Stahl, Arda Arslan, Ulrik Brandes, and Thomas Weitin. 2022. Modeling and predicting literary reception. *Journal of Computational Literary Studies*, 1(1):1–27.

Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco. 2017. Affective computing and sentiment analysis. In *A practical guide to sentiment analysis*, pages 1–10. Springer.

Jeanne S. Chall. 1947. This business of readability. *Educational Research Bulletin*, 26(1):1–13.

Jeanne S. Chall and Edgar Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books.

Matthew W. Crocker, Vera Demberg, and Elke Teich. 2016. Information Density and Linguistic Encoding (IDeaL). *KI - Künstliche Intelligenz*, 30(1):77–81.

Tess Crosbie, Tim French, and Marc Conrad. 2013. Towards a model for replicating aesthetic literary appreciation. In *Proceedings of the Fifth Workshop on Semantic Web Information Management*, SWIM '13, pages 1–4, New York, NY, USA. Association for Computing Machinery.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational Research Bulletin*, 27(1):11–28.

Wayne A. Danielson, Dominic L. Lasorsa, and Dae S. Im. 1992. Journalists and novelists: A study of diverging styles. *Journalism Quarterly*, 69(2):436–446.

Stefania Degaetano-Ortlieb and Elke Teich. 2022. Toward an optimal code for communication: The case of scientific english. *Corpus Linguistics and Linguistic Theory*, 18(1):175–207.

Alfonso Delgado-Bonal and Alexander Marshak. 2019. Approximate Entropy and Sample Entropy: A Comprehensive Tutorial. *Entropy*, 21(6):541.

Irina-Ana Drobot. 2013. Affective narratology. the emotional structure of stories. *Philologica Jassyensia*, 9(2):338.

William Dubay. 2004. *The Principles of Readability*. Impact Information.

Katherine Elkins. 2022. *The Shapes of Stories: Sentiment Analysis for Narrative*. Cambridge University Press.

Lee Fleisher, Steve Pincus, and Stanley Rosenbaum. 1993. Approximate entropy of heart rate as a correlate of postoperative ventricular dysfunction. *Anesthesiology*, 78(4):683—692.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32:221–233.

Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. 2013. Success with style: Using writing style to predict the success of novels. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1764, Seattle, Washington, USA. Association for Computational Linguistics.

Jianbo Gao, Jing Hu, and Wen-wen Tung. 2011. Facilitating Joint Chaos and Fractal Analysis of Biosignals through Nonlinear Adaptive Filtering. *PLoS ONE*, 6(9):e24331.

Jianbo Gao, Matthew L Jockers, John Laudun, and Timothy Tangherlini. 2016. A multiscale theory for the dynamical evolution of sentiment in novels. In *2016 International Conference on Behavioral, Economic and Socio-cultural Computing (BESC)*, pages 1–4. IEEE.

Craig L. Garthwaite. 2014. Demand spillovers, combative advertising, and celebrity endorsements. *American Economic Journal: Applied Economics*, 6(2):76–104.

Édouard Glissant. 1997. *Poetics of relation*. University of Michigan Press, Ann Arbor.

Qiyue Hu, Bin Liu, Mads Rosendahl Thomsen, Jianbo Gao, and Kristoffer L Nielbo. 2021. Dynamic evolution of sentiments in never let me go: Insights from multifractal theory and its implications for literary analysis. *Digital Scholarship in the Humanities*, 36(2):322–332.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

SM Mazharul Islam, Xin Luna Dong, and Gerard de Melo. 2020. Domain-specific sentiment lexicons induced from labeled documents. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6576–6587.

Swapnil Jain, Shrikant Malviya, Rohit Mishra, and Uma Shanker Tiwary. 2017. Sentiment analysis: An empirical comparative study of various machine learning approaches. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 112–121, Kolkata, India. NLP Association of India.

Matthew Jockers. 2017. Syuzhet: Extracts sentiment and sentiment-derived plot arcs from text (version 1.0. 1).

Evgeny Kim and Roman Klinger. 2018. A survey on sentiment and emotion analysis for computational literary studies. *arXiv preprint arXiv:1808.03137*.

Evgeny Kim, Sebastian Padó, and Roman Klinger. 2017. Investigating the relationship between literary genres and emotional plot development. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 17–26, Vancouver, Canada. Association for Computational Linguistics.

Courtney Kimball. 2017. Sweeney todd's dreadfuls and mass readership. *The Journal of Publishing Culture*, 7:1–12.

Irwin S. Kirsch, United States, Educational Testing Service, and National Center for Education Statistics, editors. 1993. *Adult literacy in America: a first look at the results of the National Adult Literacy Survey*, 2nd ed edition. Office of Educational Research and Improvement, U.S. Dept. of Education, Washington, D.C.

Jon P. Klancher. 1983. From "crowd" to "audience": The making of an english mass readership in the nineteenth century. *ELH*, 50(1):155–173.

Corina Koolen, Karina van Dalen-Oskam, Andreas van Cranenburgh, and Erica Nagelhout. 2020. Literary quality in the eye of the Dutch reader: The national reader survey. *Poetics*, 79:1–13.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting BERT for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 34–41, Hong Kong, China. Association for Computational Linguistics.

Dallas Liddle. 2019. Could Fiction Have an Information History? Statistical Probability and the Rise of the Novel. *Journal of Cultural Analytics*, page 22.

Suraj Maharjan, John Arevalo, Manuel Montes, Fabio A. González, and Thamar Solorio. 2017. A multi-task approach to predict likability of books. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1217–1227, Valencia, Spain. Association for Computational Linguistics.

Suraj Maharjan, Sudipta Kar, Manuel Montes, Fabio A. González, and Thamar Solorio. 2018. Letting emotions flow: Success prediction by modeling the flow of emotions in books. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Volume 2, Short Papers*, pages 259–265, New Orleans, Louisiana. Association for Computational Linguistics.

Benoit Mandelbrot. 1982. *The Fractal Geometry of Nature*. Times Books, San Francisco.

Claude Martin. 1996. Production, content, and uses of bestselling books in quebec. *Canadian Journal of Communication*, 21(4).

Harry G. McLaughlin. 1969. Smog grading: A new readability formula. *Journal of Reading*, 12(1):639–646.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Saif Mohammad and Peter Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada*, 2:1–234.

Mahdi Mohseni, Christoph Redies, and Volker Gast. 2022. Approximate entropy in canonical and non-canonical fiction. *Entropy*, 24(2):278.

Arthur K. Moore. 1964. The case for poetic obscurity. *Neophilologus*, 48(1):322–340.

Franco Moretti. 2000. The slaugtherhouse of literature. *MLQ: Modern Language Quarterly*, 61(1):207–227.

Mika V. Mäntylä, Daniel Graziotin, and Miikka Kuutila. 2018. The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. 27:16–32.

Steve Pincus. 1991. Approximate entropy (apen) as a complexity measure. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 5(1):110–117.

Steve Pincus, Igor Gladstone, and Richard Ehrenkranz. 1991. A regularity statistic for medical data analysis. *Journal of Clinical Monitoring*, 7(4):335–345.

Pontus Plavén-Sigray, Granville James Matheson, Björn Christian Schiffler, and William Hedley Thompson. 2017. Research: The readability of scientific texts is decreasing over time. *eLife*, 6:e27725.

Colin Pope. 2019. We Need to Talk About the Canon: Demographics in 'The Norton Anthology'.

Maja Popović, Sheila Castilho, Rudali Huidrom, and Anya Belz. 2022. Reproducing a Manual Evaluation of the Simplicity of Text Simplification System Outputs. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 80–85, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. 5(1):1–12.

Lucius A. Sherman. 1893. *Analytics of Literature: A Manual for the Objective Study of English Prose and Poetry*. Athenaeum Press. Ginn.

Sean Shesgreen. 2009. Canonizing the canonizer: A short history of the norton anthology of english literature. *Critical Inquiry*, 35(2):293–318.

Sanja Stajner, Richard Evans, Constantin Orasan, and Ruslan Mitkov. 2012. What can readability measures really tell us about text complexity? In *Proceedings of Workshop on natural language processing for improving textual accessibility*, pages 14–22, Istanbul, Turkey. Association for Computational Linguistics.

Tanja Säily, Arja Nurmi, Minna Palander-Collin, and Anita Auer, editors. 2017. *Exploring Future Paths for Historical Sociolinguistics*, volume 7 of *Advances in Historical Sociolinguistics*. John Benjamins Publishing Company, Amsterdam.

Joan Torruella and Ramon Capsada. 2013. Lexical statistics and tipological structures: A measure of lexical richness. *Procedia - Social and Behavioral Sciences*, 95:447–454.

Wen-wen Tung, Jianbo Gao, Jing Hu, and Lei Yang. 2011. Detecting chaos in heavy-noise environments. *Physical Review E*, 83(4).

T. Underwood. 2019. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press.

Ted Underwood and Jordan Sellers. 2016. The *Longue Durée* of Literary Prestige. *Modern Language Quarterly*, 77(3):321–344.

Andreas van Cranenburgh and Rens Bod. 2017. A data-oriented model of literary language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1228–1238, Valencia, Spain. Association for Computational Linguistics.

Melanie Walsh and Maria Antoniak. 2021. The goodreads 'classics': A computational study of readers, amazon, and crowdsourced amateur criticism. *Journal of Cultural Analytics*, 4:243–287.

Xindi Wang, Burcu Yucesoy, Onur Varol, Tina Eliassi-Rad, and Albert-László Barabási. 2019. Success in books: Predicting book sales before publication. *EPJ Data Science*, 8(1):31.

I. Westin. 2002. *Language Change in English Newspaper Editorials*. Language and computers : studies in practical linguistics. Rodopi.

Allon White, Lecturer in English Allon White, and White Allon. 1981. *The Uses of Obscurity: The Fiction of Early Modernism*. Routledge & Kegan Paul.

Marna K. Winter and Kristen O'Neill. 2022. An exploration of prevalence and usage of hi-lo texts in today's classrooms. *Reading & Writing Quarterly*, 0(0):1–15.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2020. DomBERT: Domain-oriented language model for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1725–1731, Online. Association for Computational Linguistics.

Bridging the Gap: Demonstrating the Applicability of Linguistic Analysis Tools in Digital Musicology

Sebastian Oliver Eck
Department of Musicology Weimar-Jena
University of Music Franz Liszt Weimar, Germany
sebastian.eck@hfm.uni-weimar.de

Abstract

This study introduces the novel concepts of Explicit and Implicit Musical Parameters (EMPs and IMPs) and demonstrates their application in digital musicology. Furthermore, it discusses the concept of 'musical words', that suggests representing explicit and implicit musical parameters as words or textual entities. This 'music-to-text' approach allows the application of advanced techniques and tools commonly used within the computational linguistics for the analysis of musical data, highlighting the structural parallels between music and language. Lastly, the findings of this paper not only illustrate the feasibility of this approach but also pave the way for further interdisciplinary studies and the advancement of analytical user-friendly tools that are applicable in both computational linguistics and digital musicology.

1 Introduction

In the age of digital humanities, interdisciplinary dialogue between interrelated scientific fields is becoming increasingly important. With the intentions of showing the possible benefits of interdisciplinary approaches, this paper investigates to what extent easy-to-use off-the-shelf software-tools and methods commonly used within the computational linguistics (CL) can be adopted and applied by *digital musicologists* for their own research within their respective field.

An initial assessment of the relevant scientific literature underscores the urgency and relevance of such an exploration: though the potential merits (as well the presumed deficiencies) of using computational methods in musicology have already been recognised and emphasized decades ago (Volk et al., 2011; Cook, 2005; Huron, 1999), data available on lens.org, a repository of worldwide patent and academic knowledge, reveals a significant and widening gap between the number of publications in *digital musicology* and those in *computational linguistics* (cf. Figure 1). It is reasonable to assume that this discrepancy might not only persist but potentially grow wider, as with the growing public interest and economic relevance of natural language processing and generation technologies, such as the now famous ChatGPT, computational linguistics is likely to gain even more resources, both in academia as well as the free market. However, the significance of this research paper lies in its potential to create a bridge between two seemingly disparate fields: successfully applying CL tools within the field of digital musicology could not only expand research possibilities drastically, but also facilitate a more holistic approach when researching and trying to understand the similarities between language and music.

This study is structured into four main parts: an introduction that discusses structural parallels between music and language, further introducing the novel idea of implicit and explicit musical parameters (EMPs and IMPs); a data section describing the data preparation process, encompassing the collection, tokenization and transformation of data used for this study; a methodology section, in which will be explained, how GUI-applications frequently used in computational linguistics, such as AntConc, can be adapted to analyse textual music data; and finally, a demonstration section, in which methods are presented that showcase how AntConc could be employed to address various musicological questions.

1.1 Related Work

This study positions itself in the broad field of digital musicology by explicitly discussing the concept of *musical words* to which computational linguistics (CL) tools can be applied. It establishes connections between music and language, exploring

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 248–265
December 1-3, 2023 ©2023 Association for Computational Linguistics

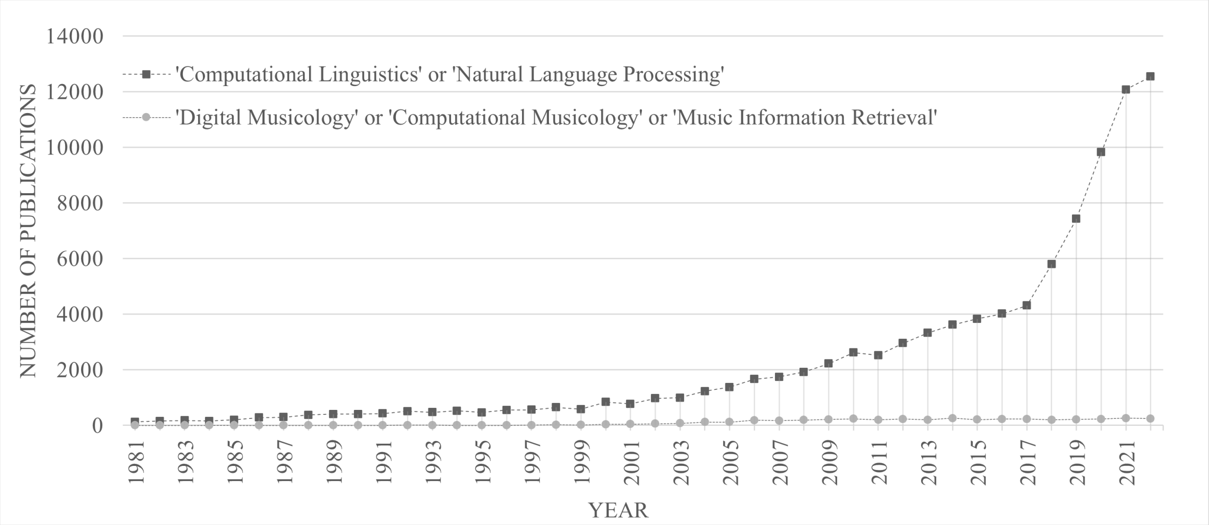


Figure 1: Scientific Publication Trends Measured by Keyword Matches in Title, Abstract, Keyword, or Field of Study (Data Source: lens.org)

their shared structural characteristics, in order to integrate CL tools into musicological research. As the work identifies a lack in studies in the application of CL tools to music data, it highlights opportunities for interdisciplinary research. This paper's findings, in particular the presented *music-to-text approach* further supports the foundational framework, set out by others (Norgaard and Römer, 2022; Wołkowicz et al., 2008), for reinterpreting monophonic musical data, e.g., folk songs, as music text, structurally comparable to written language. This paper differentiates itself by addressing theoretical limitations and philosophical interpretations of musical notation and performance, particularly through introducing the novel concepts of Explicit Musical Parameters (EMPs) and Implicit Musical Parameters (IMPs). On a final note, this study aims to further lower the barriers for interdisciplinary research by demonstrating the applicability of user-friendly GUI computational linguistics tools for musicological research.

Despite its rather low absolute quantity of publications, over the last decades research in computational musicology or, more specifically, Music Information Retrieval (MIR) has seen significant contributions in various areas. This includes work in authorship and composer recognition (Hontanilla et al., 2013; Kaliakatsos-Papakostas et al., 2010; Van Kranenburg and Backer, 2005), as well as in artist similarity recognition (Shao et al., 2008), genre recognition (Mayer and Rauber, 2011), or music recognition using 'acoustic fingerprinting' as an example (Brinkman et al., 2016).

Monophonic folk music classification, with its relatively straightforward structure, has consistently been a focal point in Music Information Retrieval (MIR) research (Huron et al., 1996). This trend continues in more recent studies (Hillewaere et al., 2014, 2009a; Taminau et al., 2009). In contrast, polyphonic music information retrieval, due to its inherent complexities, proofs to be more challenging, as evidenced in various conducted studies (Hillewaere et al., 2010, 2009b).

Additionally, MIR has made significant progress in the context of chord embeddings (Lahnala et al., 2021) and the development of chord vector representations (Madjiheurem et al., 2016).

Lastly, MIR's advancements have also extended beyond music, incorporating techniques that might seem more conventional from a linguistic viewpoint that were utilized for song lyric analysis (Mahedero et al., 2005) or classification (Fell and Sporleder, 2014).[1]

The study critically analyzes the limitations of its data and adopts both a humanities and computational perspective, enhancing the merit of its approach. This work aims to open new pathways in digital and computational musicology to pave the way for future interdisciplinary research endeavors.

[1] The author wishes to express gratitude to the peer reviewers for their insightful literature recommendations, which have significantly contributed to enriching the context and depth of this study.

1.2 Comparing the Structure of Music and Language

One key assumption of this paper is that (written) music and (written) language share enough formal similarities to make these two distinct sign systems structurally comparable and research tools mutually applicable. However, finding and defining those similarities is not an easy task: for instance, both music and spoken language follow an intrinsic logic; syntax and grammar give, to a certain degree, meanings to otherwise arbitrarily combined units, such as words, letters, or as in the case of music, musical notes. But unlike in linguistics, where many broadly available tools have already proven to rather reliably identify and give syntactic meaning to elements of, for example, a sentence, for music, in which even the definition of a phrase, a melody or the function of a chord is mostly ambiguous, such dependable instruments and methods are yet to be found and developed.[2]

One reason might be, of course, that in contrast to music, language conveys a clear, human understandable message, and therefore the correctness of the grammatical rule-set applied, can rather easily be verified by a listener familiar with the language in question. In music, the concepts of syntax and grammar seem less concrete, but equally context dependent: the defining rules of music greatly depend on factors such as the historical and cultural context in which the piece was composed, its genre, as well as the theoretical and cultural background of its composer or listener. Even though shared structural commonalities can be identified within one or across a set of several musical pieces, these, it seems, are not as strictly defined as the rules of language.

1.3 The Limitations of Sheet Music Notation

Despite this clear lack of an universally defined intrinsic musical structure, a vast number of musical parameters can still be extracted and, consequently, patterns can, presuming their existence, be identified. However, asking the question of how to extract those parameters is particularly intriguing when considering the complex nature of music.

Sheet music notation is best described as a textual, time-continuous, simplified representation of musical reality. Whereas musical reality - due to its subjective nature - still lacks a clear definition, from a purely physical perspective, music is generally understood as a series of sound events that we interpret and *understand* as music. As a consequence, any sheet music notation, digital or analogue, is by nature a mere simplification or abstraction of an indefinitely *complex musical reality*. To get back to and support the initial assumption of this paper, based on these observations it seems reasonable to argue that sheet music notation, in other words *written music*, is as much a simplified description of *performed music* as *written text* is a simplified description of *spoken language* - therefore, no representation, of musical or literary nature, will ever be able to entirely represent physical reality in its entire complexity; as a general verdict, we can presume that some information loss is *inevitable*.

This, of course, comes with difficulties, as much of the information - that had existed or will exist in the exact moment text turns into sound, is not (yet) contained in our textual source material. But this exact apparent shortcoming, on the other side, reduces the information that we need to work with to a humanly graspable, but in context of this study even more importantly, computationally manageable amount. More over, notation systems bear the potential to enrich descriptions as they give space for including information that is not inherent in the object or phenomenon that they describe (e.g., performance directions, cross-references, subdivisions etc.).

Due to these obvious limitations, music notation systems generally represent only a very limited set of musical parameters. The amount as well as the specific set chosen are usually determined by its anticipated scope of application scenarios. As the number of specific scenarios is virtually

[2] While the task of defining and finding similarities between music and language is complex, notable attempts have been made in this area. For instance, a previous, rather extensive publication (Granroth-Wilding, 2013) demonstrates the application of Combinatory Categorial Grammar (CCG) to analyze the hierarchical structure of chord sequences. This approach introduces a formal language, similar to first-order predicate logic, to express the tonal harmonic relationships between chords, serving as a mechanism to map unstructured chord sequences into structured analyses. Additionally, a subsequent study (Granroth-Wilding and Steedman, 2014) successfully showed the effectiveness of applying machine learning techniques to the identification of musical grammar. It describes the use of a formal grammar of jazz chord sequences, combined with statistical modelling techniques, for parsing musical structure, demonstrating that these NLP-adapted statistical techniques can be profitably applied to the analysis of generally ambiguous harmonic structure in music. Further, as a side note, James R. Meehan's work (Meehan, 1979) is notable as it provides a rather historic, however interesting comparison between language and music originating from the early days of Artificial Intelligence (AI).

unlimited, over the past decades, parallel to the development of various computational methods for music analysis, various digital formats for storing music (related) information have already been, are expected to be invented or further developed. For a long time, some of the most commonly used music notation systems have been Standard MIDI (.mid) (Loy, 1985)[3] as well as the much younger MusicXML (.mxl/.musicxml) (Good, 2001),[4] both formats focusing on a rather practical aspect of storing music information for playback, or as in case of musicXML for representation in music notation software. Nowadays, the Music Encoding Initiative's schema MEI (Roland, 2000; Hankinson et al., 2011),[5] has established itself as the gold standard for academic music notation. Analogue to the Text Encoding Initiative's format TEI (Aguera et al., 1987),[6] MEI was invented in particular with the intentions to standardise music encoding for scientific and archival use.[7]

1.4 Introducing the Concept of Explicit and Implicit Musical Parameters

When trying to retrieve musical parameters from any form of textual music representation, analogue or digital, it seems reasonable to differentiate between explicit and implicit musical parameters contained within the given source material. As a consequence, the following classification is proposed:

In the context of Music Information Retrieval (MIR) the term explicit musical parameter (EMP) must refer to a musical attribute that is explicitly notated or indicated in the given sheet music notation. The term implicit musical parameter (IMP), on the other hand, should refer to a musical attribute that is not directly stated but can be inferred or deduced, i.e., by comparison or calculation.

EMPs are musical parameters that are explicitly defined and communicated within the given musical notation system. As in .xml-notation, this information would be encoded using specific, predefined symbols and coding conventions. Within this code, those symbols and markings store absolute information, such as pitches, durations, dynamics, tempo markings, articulations, as they were encoded by the annotator. However, given the complex nature of music (cf. Chapter 1.3), any notation system can represent/contain only a finite number of EMPs, which, as a side note, further underlines the inevitable limitations of musical notation as a static representation of musical reality.

IMPs, on the other side, represent relational or contextual information. They aren't directly stated but are deduced by observing changes in one or more EMPs or IMPs across multiple consecutive musical events. For instance, in the context of the aforementioned .xml notation format, a musical interval would be considered an implicit musical parameter, as it is not explicitly notated but can be inferred from the difference in pitch between two compared notes.

As the number of performable calculations is, at least in theory, indefinite, any form of sheet music notation, whether digital or analogue, theoretically contains an equally infinite number of IMPs; of which, of course, not all are equally relevant for music analysis or related research.

Differentiating between EMPs and IMPs helps to better understand and handle complex musical information: EMPs, being directly indicated, form the basic layer of information that is relatively easy to access and process; IMPs, on the other hand, though not directly represented, bring additional levels of complexity which can be calculated and utilized when needed, and ignored when not. This can lead to better and more efficient musical parameter extraction methods, as well as rank and classify musical data representations according to the number of EMPs contained.

1.5 The Use of CL Tools for Music Information Retrieval

With this understanding of explicit and implicit musical parameters well established, we shall now turn to the question of how to navigate and make sense of this vast, sheer endless set of complex information. Luckily, the need to manage and extract patterns from a substantial and often ambiguous dataset is not unique to music: within the field of computational linguistics (CL) many of such explorative tools have already been developed and adopted to a variety of different use-cases. For instance, in the CL those patterns are usually found within repeating combinations of words, letters, or other linguistic elements, expressed as n-grams,

[3] https://www.midi.org/

[4] https://www.musicxml.com/

[5] https://music-encoding.org/

[6] https://tei-c.org/

[7] For a detailed discussion on the challenges and complexities of digitally encoding music notation compared to text encoding, particularly through the Music Encoding Initiative's (MEI) and Text Encoding Initiative's (TEI) respective formats, see (Teich Geertinger, 2021)

which are contiguous sequences of 'n' items from a given sample of text. Unsurprisingly, utilizing n-gram searches analogously in digital musicology research to identify recurring patterns in sequences of musical elements has already been well established, as exemplified in various studies (cf. section 1.1).

Those linguistic elements, seen as the smallest distinguishable units, are typically referred to as tokens. By applying the concept of tokens to music, these then newly created *musical tokens*, which could encompass a variety of explicit and implicit parameters such as individual pitches, durations, or as an example for relational information, horizontal (melodic) intervals, can, as this paper will show in its last section, facilitate the application of CL tools on musical source material, and therefore to a certain degree, close the gap between the computational linguistics and music information retrieval.

2 Data

The second section of this paper will focus on data collecting and preparation. In this process several tools where utilized to convert musical source material into textual data, as necessitated by the tool AntConc used for corpus analysis in sections 3 and 4 of this study (cf. Figure 2).

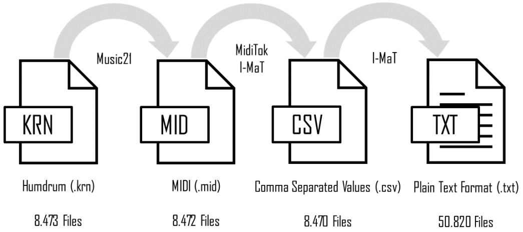


Figure 2: Musical Data Conversion and Custom Database Creation: From Kern Score to MIDI to CSV to Textual Data

The complete process can be divided into two sub-sections: *Corpus Creation* as well as *Tokenization and Database Creation*.

2.1 Corpus Creation

A selection of roughly 8.000 music files that belong to the Essen Folksong Database (Schaffrath, 1997) were chosen as the source material for the study at hand. With more than "20.000 songs and instrumental melodies, mostly from Germany, Poland and China, with minor collections from other (mostly European) countries" (Dahlig, n.d.), the Essen Folksong Database offers an extensive set of diverse monophonic musical pieces (cf. Figure 3). As of today, 8.473 pieces within The Essen Folksong Database are available as Humdrum data translations (Schaffrath and Huron, 1995; Sapp, 2005). Referenced from its associated website,[8] the **kern (.krn) based Humdrum file format is best described as a "text-based description for musical scores, and its primary purpose is for computational musical analysis using the Humdrum Toolkit." The Humdrum format is capable of containing metadata as well as basic music information for each individual musical event, such as pitch, duration, key signature, tempo, meter and others (cf. Figure 5). However, for compatibility with the chosen tool for tokenization, those files needed to be converted to a more universally accepted format: the aforementioned Musical Instrument Digital Interface (MIDI, .mid) file format for music representation. The conversion was accomplished by using the music21 Python framework (Cuthbert and Ariza, 2010), developed by the Massachusetts Institute of Technology (MIT), that offers a variety of tools for handling and manipulating music data.

2.2 Tokenization and Database Creation

After successfully converting nearly all of the available 8.473 .krn-files, with one exception, the resulting 8.472 .mid-files were prepared for further processing. This next step involved the extraction of certain explicit musical parameters for each successive musical event (also commonly referred to as musical *note*). This process, known as tokenization, involved, in the scope of this study, breaking down each musical piece and its entire melodic line into its fundamental components or, as termed earlier, *musical tokens*.

2.2.1 Tokenization: MidiTok

In this study, these *musical tokens* were created utilizing MidiTok (Fradet et al., 2021), an open-source Python package for MIDI file tokenization.[9] MidiTok offers a variety of ten different tokenizers, each of which uses a different pattern to combine several extracted explicit musical parameters (EMPs) into distinct musical tokens. The tokens

[8] http://kern.humdrum.org/help/tour/

[9] https://miditok.readthedocs.io/en/v2.1.7/index.html

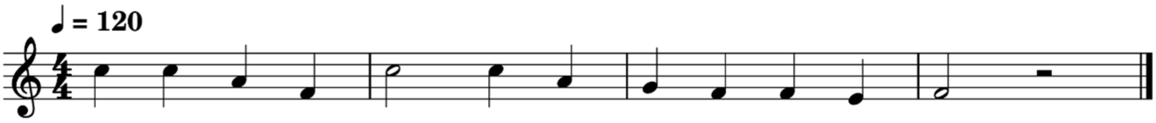


Figure 3: Score Representation of the Folk Song "Nun schürz dich, Gretlein"; signature: deut0781

are stored in either one-dimensional lists (cf. Figure 6) or two-dimensional nested lists (cf. Figure 7), with the latter format proving useful for data transformation as it allows each musical token to be allocated to a separate array. Those arrays can easily be extracted and stored within a more universal table-like data structure for further processing using computational methods.

Tokenizer	Success Rate (%)	Structure
MIDILike	99.976	1D
MMM	99.976	1D
REMI	99.976	1D
Structured	99.976	1D
TSD	99.976	1D
CPWord	99.976	2D
MuMIDI	99.976	2D
OctupleMono	99.976	2D
Octuple	88.302	2D
REMIPlus	88.302	1D

Table 1: Success rates and structural dimensions of MidiTok tokenizers in processing 8472 .mid-files (sorted by Success Rate (%) in descending order).

Although each of MidiTok's ten tokenizers was applied on the data set equally, not all tokenizers performed with equal reliability. Specifically, the **Octuple** (Zeng et al., 2021) and **REMIPlus** (von Rütte et al., 2022) tokenizers failed to tokenize 991 of the in total 8472 .midi-files (11.697%) (cf. Table 1). Consequently, only the remaining eight tokenizers (1D: **MIDILike** (Oore et al., 2018), **MMM** (Ens and Pasquier, 2020)), **REMI** (Huang and Yang, 2020), **Structured** (Hadjeres and Crestel, 2021), **TSD** (Fradet et al., 2023); 2D: **CPWord** (Hsiao et al., 2021), **MuMIDI** (Ren et al., 2020) and **OctupleMono**) were considered for further study. In the end, **OctupleMono**[10] was chosen

1. for its *high reliability*, failing only two tokenizations out of 8472 files, and

2. its *two-dimensional structure*, which enables presorted parameters and efficient data handling. As shown in Figure 7, this structure is unique to OctupleMono as well as other two-dimensional tokenizers, offering a more organized approach compared to one-dimensional tokenizers due to their grouped parameters or tokens.

In a last data transformation process, the tokenized data of each of the 8470 successfully processed files were compiled into a single .csv file, with the first column being the individual file names, facilitating further data manipulation in subsequent steps. For a visual representation of this compiled data, refer to Table 4.

2.2.2 Data Refinement

The data refinement involved several steps. Initially, the 'Duration' values were converted into numeric values, making them easier to read, eventually calculate on and compare. Following this, a relational implicit musical parameter was calculated: The 'PitchDifferenceToNextPitch' parameter would later (cf. section 4) allow for a horizontal (melodic) interval search while also including its transpositions. Another refinement step involved the removal of redundant prefixes following the pattern "prefix_". This streamlined the data and reduced its overall size (cf. Table 5).

2.2.3 Data Extraction

The last phase of data preparation involved creating individualized data files for each piece and parameter. Essentially, for each music piece, every musical parameter tied to its filename was separated out.

These separated data sets were then saved as individual .txt files within a dedicated folder (cf. Figure 8). It is important to note that each of these .txt-files contained only a single type of extracted parameter (cf. Figure 4), making the data readable

[10] The OctupleMono tokenizer is constructed similarly to the Octuple tokenizer (Zeng et al., 2021). The difference is the exclusion of the 'Program token' in OctupleMono, making it specifically suitable for the tokenization of monophonic music files, which consist of a single track.

by the concordance and analysis tool AntConc used in sections 3 and 4.

```
0.0 -3.0 -4.0 7.0 0.0 -3.0 -2.0 -2.0 0.0 -1.0 1.0 nan
```

Figure 4: Extracted Relational Token "PitchDifferenceToNextPitch" of the folk song "Nun schürz dich, Gretlein"; signature: deut0781

With these steps, the necessary custom database was created and prepared for the next phase of the study, ensuring a solid and easily accessible data foundation for the subsequent corpus analysis.[11]

Note: All the data conversion and tokenization steps outlined in this section have been integrated into the Interactive Music Analysis Tool (I-MaT) python package (Eck, 2023). I-MaT, complete with its robust functionalities for corpus creation, database creation, tokenization, as well as music analysis is openly available for access and usage via GitHub.[12]

3 Methodology

In the following section, the methodological approach, used for demonstrating the applicability of well-established CL tools for answering questions emanating from the field of musicology will be presented.[13]

3.1 The Text Concordance and Analysis Tool AntConc

The tool of choice was the freeware corpus analysis toolkit AntConc (Anthony, 2004), which can be described as an easy-to-use tool originally developed for concordancing and text analysis. AntConc uses several tools that include pattern search, pattern distribution analysis and similar functions for analysing individual text files or conducting corpus studies. AntConc 4.2.0[14] offers these functions via the Keyword in Context (KWIC) tool, the Plot tool, and other tools such as Cluster, N-Gram, Collocate, Wordlist, and Keyword-List.

3.2 Utilizing AntConc for statistical analysis in the context of MIR

In a first step, AntConc's in-built Corpus Manager was used to create a custom database. For this task, the folder "PitchDifferenceToNextPitch" created earlier served as the source, containing all the necessary .txt-files to be included in the custom corpus. A critical aspect to consider during the corpus creation was to reconfigure the 'token_definition' setting: to ensure an accurate computation of types and tokens, this parameter was adjusted to include all numerical values and ".na" (value: [-0123456789.na]). This step was essential as the data in use primarily consists of numerical values but also possible "nan" [= no subsequent interval available] entries (cf. Chapter 4.2).

Utilizing AntConc for statistical analysis in the context of MIR demanded a conceptual re-framing: numerical values, which represent musical parameters, are treated as words. In theory, any desired number of numerical values, i.e., numerical representations of musical parameters, extracted from one and the same musical event (e.g., pitch, duration etc.) can be combined. A higher number would result in more complex, but also more exact representations of musical events.

For simplicity, in this study only one musical parameter was extracted from each musical event. These single-element word-sequences are separated by whitespace characters, imitating the structure of a sentence (cf. Figure 4). This unconventional approach is necessitated by the nature of the tool, which is designed primarily for text analysis. In short, for AntConc to correctly interpret the musical text data at hand, numbers had to be reinterpreted as numerical representations of musical events, as words ultimately forming sentences.

Though this approach might seem unusual, it allowed, as the final section will show, to nearly fully engage the capabilities of AntConc in a musicological context. This 'numerical-to-textual' approach made it possible, as the following examples show, to perform n-gram pattern searches, conduct cluster analyses on and plot pattern distributions on these newly created 'musical words' and 'sentences'.

[11] All the data (.mid-conversions, tokenizations, cleaned and enhanced databases (.csv-files), and extracted parameters (.txt-files) used in this study are freely available on figshare.com for further research and exploration: 'https://figshare.com/s/03179521a56c88bfac63'.

[12] https://github.com/sebastian-eck/I-MaT

[13] During the review process, the author has been made aware of a similar approach recently attempted (Norgaard and Römer, 2022), but as the work is not readily accessible, this study still offers unique contributions to digital musicology with its distinct use case and theoretical framework.

[14] https://www.laurenceanthony.net/software/antconc/.

4 Demonstration

In this last section, three distinct ways to utilize AntConc for music analysis will be demonstrated. Each of these demonstrations will focus on a particular aspect of music analysis:

1. **melodic pattern search** (N-Gram-Tool)

2. **end phrase pattern search** (Cluster-Tool)

3. **pattern distribution** (Plot-Tool)

These analyses will be performed on the custom database, created within AntConc earlier in section 2.2. Numerical representations of musical events will be treated as textual entities or ’musical words’, allowing AntConc to be used as an user-friendly tool for conducting various music analyses.

Note: As the pattern search will be performed not on absolute pitch values, but on values extracted from the relational parameter “PitchDifferenceToNextPitch”, patterns will be identified across melodic transpositions.

4.1 N-Gram-Tool

Identifying Common Melodic Sequences: One musicological question that can be addressed using the n-gram pattern search feature of AntConc is identifying the most common melodic sequences within a corpus of music. After interpreting sequences of relative pitch intervals, or ’musical words’, as n-grams, the pattern search function revealed several sequences as the most prominent (cf. Table 2).

Type	Rank	Freq	Range
`0.0 0.0 0.0 0.0 0.0`	1	1210	441
`-2.0 -2.0 -1.0 -2.0 -2.0`	2	894	683
`2.0 -2.0 -2.0 -1.0 -2.0`	3	846	637
`2.0 1.0 -1.0 -2.0 -2.0`	4	555	436
`-2.0 -1.0 -2.0 -2.0 0.0`	5	543	455
`-2.0 -1.0 -2.0 -2.0 2.0`	6	519	413
`2.0 2.0 -2.0 -2.0 -1.0`	7	499	363
`3.0 -2.0 -1.0 -2.0 -2.0`	8	494	373
`1.0 2.0 2.0 -2.0 -2.0`	9	484	353
`2.0 2.0 1.0 -1.0 -2.0`	10	469	352

Table 2: Results (Excerpt) of a N-Gram-Search (n = 5) Performed on the Custom Database

These patterns can be grouped as:

1. **Repetitive Sequences:** {0.0 0.0 0.0 0.0 0.0}

2. **Descending Interval Patterns:** {-2.0 -2.0 -1.0 -2.0 -2.0} and {-2.0 -1.0 -2.0 -2.0 0.0}

3. **Combination of Ascending and Descending Intervals:** {2.0 -2.0 -2.0 -1.0 -2.0} and {2.0 1.0 -1.0 -2.0 -2.0}

4. **Alternating Interval Structures:** {2.0 2.0 -2.0 -2.0 -1.0} and {3.0 -2.0 -1.0 -2.0 -2.0}

Conclusions: ”Regional and Cultural Influences” - The recurrence of certain patterns, such as repetitive sequences or descending interval patterns, might indicate patterns commonly used in melody construction. Given the diversity of the Essen Folksong Collection, a n-gram-search that groups results by geographic information (as indicated in the individual file names, e.g., ”steier09...”, cf. Table 9) could reveal specific patterns influenced by regional or cultural traditions, revealing patterns more prevalent in Germanic folk songs compared to those from other regions, in relation to their total number.

Note: Here, as well as in the following examples, melodic intervals are represented as numeric values, each whole number representing a semitone step. Negative values represent descending, positive values ascending intervals. Zero-values, such as in the most prominent interval-series (cf. Table 2), represent repetitions of the same pitch.

4.2 Cluster-Tool

Identifying Common End Phrase Patterns: By employing AntConc’s pattern search capabilities, musicologists can identify common end phrase patterns in a corpus of music. This can be achieved by searching for recurring sequences at the end of musical phrases.

Cluster	Rank	Freq	Range
`-2.0 -1.0 -2.0 -2.0 nan`	1	423	423
`-2.0 -2.0 2.0 -2.0 nan`	2	160	160
`0.0 -2.0 0.0 -2.0 nan`	3	157	157
`-2.0 -2.0 -1.0 1.0 nan`	4	132	132
`2.0 2.0 -2.0 -2.0 nan`	5	119	119
`-2.0 2.0 -2.0 -2.0 nan`	6	98	98
`-2.0 -2.0 -1.0 -2.0 nan`	7	95	95
`-2.0 -2.0 -3.0 -2.0 nan`	8	83	83
`1.0 -1.0 -2.0 -2.0 nan`	9	82	82
`-1.0 -2.0 0.0 -2.0 nan`	10	80	80

Table 3: Results (Excerpt) of a Cluster Search (n = 5; Search Term Position = On Right; Search Term = nan) Performed on the Custom Database

Based on the table results, we can group the common end phrase patterns as follows:

1. **Descending Endings:** The most frequent pattern, {-2.0 -1.0 -2.0 -2.0 nan}, can be interpreted as a part of a descending major scale, corresponding to, e.g., ”G F E D C” in a major key. This pattern, as well as others like {-2.0 -2.0 2.0 -2.0 nan}, suggests a common use of descending intervals at the end of phrases. The prevalence of such patterns, particularly the descending major scale, might not sound surprising to many listeners due to its widespread use in traditional folk music. But more interestingly, it reflects an expected stylistic or structural preference in folk song compositions for simplicity, aligning with the perceived conventional or rather ’simple’ nature of folk music melodies.

2. **Static and Minor Movements:** Patterns such as {0.0 -2.0 0.0 -2.0 nan} and {-2.0 -2.0 -1.0 1.0 nan} demonstrate either static (repeated pitches) or minor interval movements. These may reflect a simplicity and compactness in phrase endings commonly found in folk songs.

3. **Ascending and Mixed Intervals:** Patterns like {2.0 2.0 -2.0 -2.0 nan} and {1.0 -1.0 -2.0 -2.0 nan} include a mix of ascending and descending movements. This variety might represent a richer melodic closure in some folk songs.

Conclusions: The Cluster-Tool analysis provides insights into common phrase-ending techniques in the Essen Folksong Database. These patterns offer a glimpse into the melodic structures and stylistic tendencies in folk song compositions, particularly in how phrases are conventionally concluded. When considering geographic parameters (c.f. section 4.1), the variety and frequency of these patterns can also reflect regional or cultural influences in folk song traditions, contributing to the understanding of musicological characteristics within this genre.

Note: Here, the string ”nan” indicates the presence of end-notes. This particular string acted as a placeholder when the relational token “PitchDifferenceToNextPitch” was calculated, but no subsequent pitch value could be identified. In short, ”nan” is a marker for situations where the calculation could not continue due to the absence of a following pitch. Since this calculation was performed on sequential pitch values only, rests within the file will not be labeled. Furthermore, we can assume that the automatic inclusion of a ”start” string at the beginning of each .txt-file would effortlessly allow for an analogue ’start pattern search’.

4.3 Plot-Tool

Visualizing Pattern Distributions Across a Musical Corpus: A third application could involve using AntConc’s plotting function to visualize the distribution of specific melodic patterns across a musical corpus. For instance, a musicologist could search for a particular melodic sequence, interval pattern, or, assuming a suitable custom corpus has been created, a rhythmic motif. The results can then be displayed by using the plotting function to visually map out where and how frequently these patterns occur across different pieces (cf. Figure 9). Visualizing pattern distributions seems particularly useful within polyphonic music, such as imitative fugues, or isorhythmic motets.

5 Conclusion

This study has successfully demonstrated the practical applicability of well-established computational linguistics (CL) tools, such as AntConc, in the the field of digital musicology. The successful reinterpretation of tokenized music data as ’musical words’ outlined in sections 2 and 3 has not only indicated the existence of certain inherent structural/formal similarities between language and music; it also unveiled the potential benefits of developing new analytical tools and methods that can be used both within the fields of linguistics and musicology.

A notable area for future code development lies in the optimization of the tokenization and parameter extraction processes. Integrating more versatile tools like music21 into the tokenization process could address the current limitations encountered with MidiTok, particularly its restriction to the rather unreliable .mid-file format. Utilizing music21’s comprehensive capabilities would enable the processing of a broader spectrum of music file formats, enhancing the methodology’s versatility as well as its robustness by further reducing its dependency on third-party python packages.

Lastly, AntConc was mainly used for exemplary reasons within this study. The incorporation of advanced Natural Language Processing (NLP) packages, such as NLTK or SpaCy, directly into the aforementioned Interactive Music Analysis Tool

(I-MaT), as outlined towards the end of section 2.2, presents a significant opportunity. This integration would enable more sophisticated analysis capabilities, allowing for the calculation, exportation, and visualization of results within a unified platform. Such an approach could make the analysis process more user-friendly and accessible, particularly for scholars and students who are new to the field of digital musicology. It would further lower the barriers to entry in this interdisciplinary field, enhancing the appeal and reach of digital musicological studies.

We can assume that this research not only makes a step towards bridging the gap between computational linguistics and musicology but also lays the groundwork for a more integrated and holistic approach to the analysis of music and language.

References

Helen Aguera et al. 1987. The preparation of text encoding guidelines. Closing Statement of the Vassar Planning Conference.

Laurence Anthony. 2004. Antconc: A learner and classroom friendly, multi-platform corpus analysis toolkit. *proceedings of IWLeL*, pages 7–13.

Cors Brinkman, Manolis Fragkiadakis, and Xander Bos. 2016. Online music recognition: the echoprint system.

Nicholas Cook. 2005. Towards the complete musicologist. In *Proceedings of the Sixth International Conference on Music Information Retrieval (ISMIR 2005)*.

Michael Scott Cuthbert and Christopher Ariza. 2010. music21: A toolkit for computer-aided musicology and symbolic music data. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, pages 637–642, Utrecht, Netherlands. Version: Author's final manuscript.

Ewa Dahlig. n.d. Essen associative code and folksong database. `http://www.esac-data.org/`. Accessed on 2023/07/28.

Sebastian Oliver Eck. 2023. Interactive music analysis tool (i-mat). In *Proceedings of the JADH2023 conference*, pages 20–23. Japanese Association for Digital Humanities.

Jeff Ens and Philippe Pasquier. 2020. Mmm : Exploring conditional multi-track music generation with the transformer.

Michael Fell and Caroline Sporleder. 2014. Lyrics-based analysis and classification of music. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 620–631.

Nathan Fradet, Jean-Pierre Briot, Fabien Chhel, Amal El Fallah Seghrouchni, and Nicolas Gutowski. 2021. MidiTok: A python package for MIDI file tokenization. In *Extended Abstracts for the Late-Breaking Demo Session of the 22nd International Society for Music Information Retrieval Conference (ISMIR 2021)*.

Nathan Fradet, Jean-Pierre Briot, Fabien Chhel, Amal El Fallah Seghrouchni, and Nicolas Gutowski. 2023. Byte pair encoding for symbolic music.

Michael Good. 2001. Musicxml: An internet-friendly format for sheet music. In *Proceedings of the XML 2001 Conference*.

Mark Granroth-Wilding and Mark Steedman. 2014. A robust parser-interpreter for jazz chord sequences. *Journal of New Music Research*, 43(4):355–374.

Mark Thomas Granroth-Wilding. 2013. Harmonic analysis of music using combinatory categorial grammar.

Gaëtan Hadjeres and Léopold Crestel. 2021. The piano inpainting application.

Andrew Hankinson, Perry Roland, and Ichiro Fujinaga. 2011. The music encoding initiative as a document-encoding framework. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, pages 293–298.

Ruben Hillewaere, Bernard Manderick, and Darrell Conklin. 2009a. Global feature versus event models for folk song classification. In *ISMIR*, volume 2009, page 10th.

Ruben Hillewaere, Bernard Manderick, and Darrell Conklin. 2009b. Melodic models for polyphonic music classification. In *Second International Workshop on Machine Learning and Music*.

Ruben Hillewaere, Bernard Manderick, and Darrell Conklin. 2010. String quartet classification with monophonic models. In *ISMIR*, pages 537–542.

Ruben Hillewaere, Bernard Manderick, and Darrell Conklin. 2014. Alignment methods for folk tune classification. In *Data analysis, machine learning and knowledge discovery*, pages 369–377. Springer.

María Hontanilla, Carlos Pérez-Sancho, and Jose M Inesta. 2013. Modeling musical style with language models for composer recognition. In *Pattern Recognition and Image Analysis: 6th Iberian Conference, IbPRIA 2013, Funchal, Madeira, Portugal, June 5-7, 2013. Proceedings 6*, pages 740–748. Springer.

Wen-Yi Hsiao, Jen-Yu Liu, Yin-Cheng Yeh, and Yi-Hsuan Yang. 2021. Compound word transformer: Learning to compose full-song music over dynamic directed hypergraphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):178–186.

Yu-Siang Huang and Yi-Hsuan Yang. 2020. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1180–1188, New York, NY, USA. Association for Computing Machinery.

David Huron. 1999. The new empiricism: Systematic musicology in a postmodern age. *Berkeley, University of California*, page 2.

David Huron et al. 1996. The melodic arch in western folksongs. *Computing in Musicology*, 10:3–23.

Maximos A Kaliakatsos-Papakostas, Michael G Epitropakis, and Michael N Vrahatis. 2010. Musical composer identification through probabilistic and feedforward neural networks. In *European Conference on the Applications of Evolutionary Computation*, pages 411–420. Springer.

Allison Lahnala, Gauri Kambhatla, Jiajun Peng, Matthew Whitehead, Gillian Minnehan, Eric Guldan, Jonathan K Kummerfeld, Anıl Çamcı, and Rada Mihalcea. 2021. Chord embeddings: Analyzing what they capture and their role for next chord prediction and artist attribute prediction. In *Artificial Intelligence in Music, Sound, Art and Design: 10th International Conference, EvoMUSART 2021, Held as Part of EvoStar 2021, Virtual Event, April 7–9, 2021, Proceedings 10*, pages 171–186. Springer.

Gareth Loy. 1985. Musicians make a standard: The midi phenomenon. *Computer Music Journal*, 9(4):8–26.

Sephora Madjiheurem, Lizhen Qu, and Christian Walder. 2016. Chord2vec: Learning musical chord embeddings. In *Proceedings of the constructive machine learning workshop at 30th conference on neural information processing systems (NIPS2016), Barcelona, Spain*.

Jose P. G. Mahedero, Álvaro Martínez, Pedro Cano, Markus Koppenberger, and Fabien Gouyon. 2005. Natural language processing of lyrics. In *Proceedings of the 13th annual ACM international conference on Multimedia (MULTIMEDIA '05)*, pages 475–478, New York, NY, USA. Association for Computing Machinery.

Rudolf Mayer and Andreas Rauber. 2011. Musical genre classification by ensembles of audio and lyrics features. In *Proceedings of international conference on music information retrieval*, pages 675–680.

James R Meehan. 1979. An artificial intelligence approach to tonal music theory. In *Proceedings of the 1979 annual conference*, pages 116–120.

Martin Norgaard and Ute Römer. 2022. Patterns in music: How linguistic corpus analysis tools can be used to illuminate central aspects of jazz improvisation. *Jazz Education in Research and Practice*, 3(1):3–26.

Sageev Oore, Ian Simon, Sander Dieleman, Douglas Eck, and Karen Simonyan. 2018. This time with feeling: Learning expressive musical performance. *Neural Computing and Applications*, 32:955–967.

Yi Ren, Jinzheng He, Xu Tan, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020. Popmag: Pop music accompaniment generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, page 1198–1206. Association for Computing Machinery.

Perry Roland. 2000. Xml4mir: Extensible markup language for music information retrieval. In *Proceedings of the 1st International Society for Music Information Retrieval Conference (ISMIR 2001)*.

Dimitri von Rütte, Luca Biggio, Yannic Kilcher, and Thomas Hofmann. 2022. Figaro: Generating symbolic music with fine-grained artistic control.

Craig Stuart Sapp. 2005. Online database of scores in the humdrum file format. In *Proceedings of the ISMIR*, pages 664–665.

Helmut Schaffrath. 1997. The essen associative code: a code for folksong analysis. In Eleanor Selfridge-Field, editor, *Beyond MIDI: the handbook of musical codes*, pages 343–361. MIT Press, Cambridge, Massachusetts.

Helmut Schaffrath and David Huron. 1995. The essen folksong collection in kern format. `http://kern.ccarh.org/browse?l=essen`. Accessed on 2023/07/28.

Bo Shao, Tao Li, and Mitsunori Ogihara. 2008. Quantify music artist similarity based on style and mood. In *Proceedings of the 10th ACM workshop on web information and data management*, pages 119–124.

Jonatan Taminau, Ruben Hillewaere, Stijn Meganck, Darrell Conklin, Ann Nowé, and Bernard Manderick. 2009. Descriptive subgroup mining of folk music. In *2nd International Workshop on Machine Learning and Music (MML 2009), Bled, Slovenia*.

Axel Teich Geertinger. 2021. Digital Encoding of Music Notation with MEI. In Margrethe Støkken Bue and Annika Rockenberger, editors, *Notated Music in the Digital Sphere. Possibilities and Limitations*, volume 15 of *Nota bene – Studies from the National Library of Norway*, page 35–56. National Library of Norway, Oslo.

Peter Van Kranenburg and Eric Backer. 2005. Musical style recognition—a quantitative approach. In *Handbook of pattern recognition and computer vision*, pages 583–600. World Scientific.

Anja Volk, Frans Wiering, and Peter Van Kranenburg. 2011. Unfolding the potential of computational musicology. In *Proceedings of the 13th International Conference on Informatics and Semiotics in Organisations*, pages 137–144. Fryske Akademy.

Jacek Wołkowicz, Zbigniew Kulka, and Vlado Kešelj. 2008. N-gram-based approach to composer recognition. *Archives of Acoustics*, 33(1):43–55.

Mingliang Zeng, Xu Tan, Rui Wang, Zeqian Ju, Tao Qin, and Tie-Yan Liu. 2021. MusicBERT: Symbolic music understanding with large-scale pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 791–800, Online. Association for Computational Linguistics.

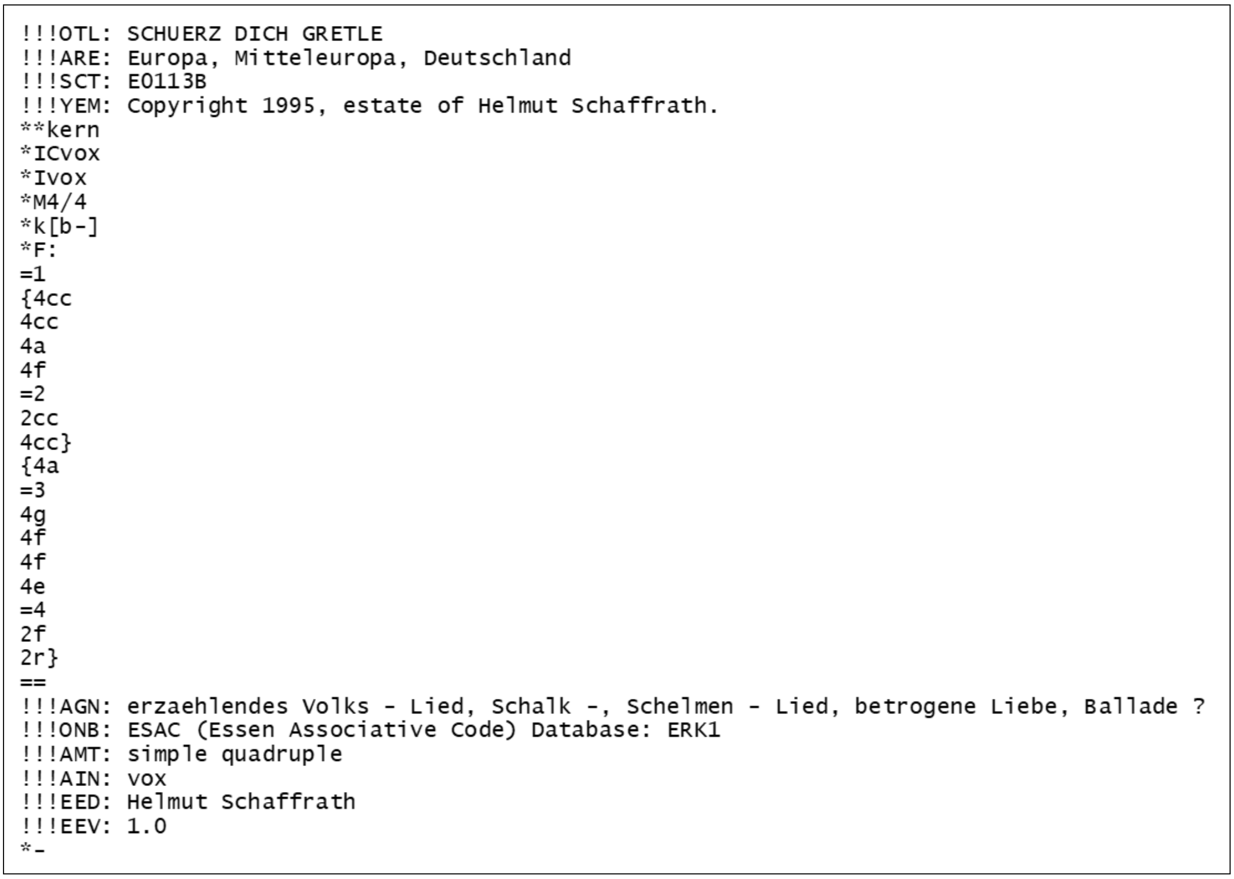
Appendices

filename	Pitch	Velocity	Duration	Position	Bar
deut0781	Pitch_72	Velocity_91	Duration_1.0.16	Position_0	Bar_0
deut0781	Pitch_72	Velocity_91	Duration_1.0.16	Position_16	Bar_0
deut0781	Pitch_69	Velocity_91	Duration_1.0.16	Position_32	Bar_0
deut0781	Pitch_65	Velocity_91	Duration_1.0.16	Position_48	Bar_0
deut0781	Pitch_72	Velocity_91	Duration_2.0.16	Position_0	Bar_1
deut0781	Pitch_72	Velocity_91	Duration_1.0.16	Position_32	Bar_1
deut0781	Pitch_69	Velocity_91	Duration_1.0.16	Position_48	Bar_1
deut0781	Pitch_67	Velocity_91	Duration_1.0.16	Position_0	Bar_2
deut0781	Pitch_65	Velocity_91	Duration_1.0.16	Position_16	Bar_2
deut0781	Pitch_65	Velocity_91	Duration_1.0.16	Position_32	Bar_2
deut0781	Pitch_64	Velocity_91	Duration_1.0.16	Position_48	Bar_2
deut0781	Pitch_65	Velocity_91	Duration_2.0.16	Position_0	Bar_3

Table 4: Database (OctupleMono) Representation of the Folk Song ”Nun schürz dich, Gretlein”; Signature: deut0781

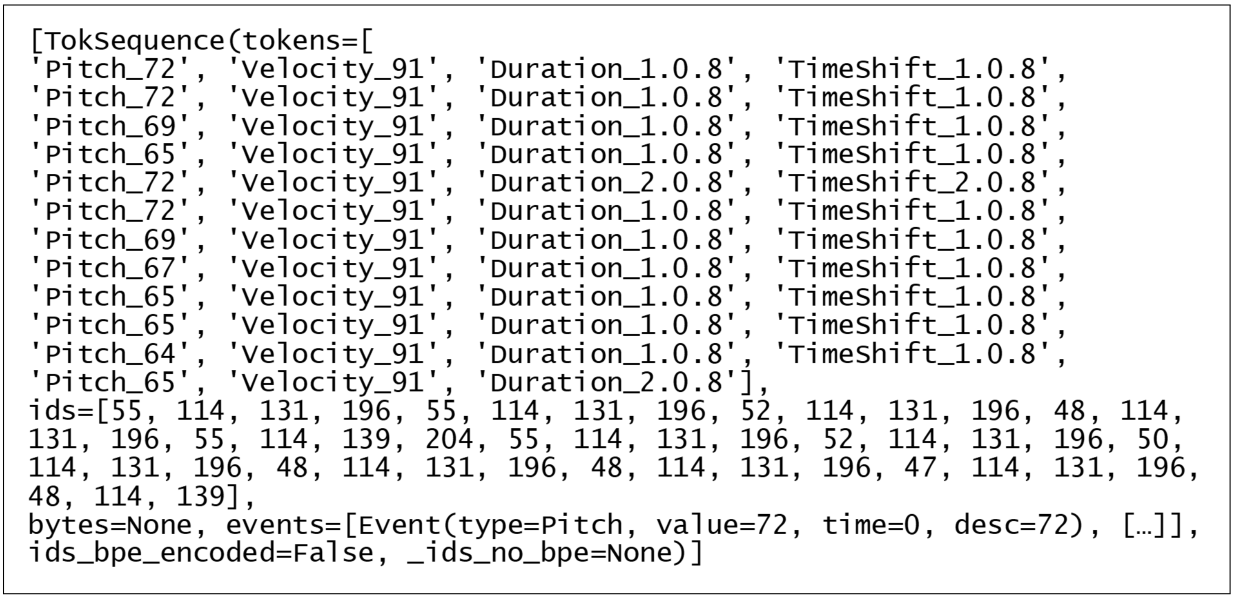
filename	Pitch	Velocity	Duration	Position	Bar	PitchDifferenceToNextPitch
deut0781	72	91	1	0	0	0
deut0781	72	91	1	16	0	-3
deut0781	69	91	1	32	0	-4
deut0781	65	91	1	48	0	7
deut0781	72	91	2	0	1	0
deut0781	72	91	1	32	1	-3
deut0781	69	91	1	48	1	-2
deut0781	67	91	1	0	2	-2
deut0781	65	91	1	16	2	0
deut0781	65	91	1	32	2	-1
deut0781	64	91	1	48	2	1
deut0781	65	91	2	0	3	nan

Table 5: Refined Database (OctupleMono) Representation of the Folk Song ”Nun schürz dich, Gretlein”; Signature: deut0781



```
!!!OTL: SCHUERZ DICH GRETLE
!!!ARE: Europa, Mitteleuropa, Deutschland
!!!SCT: E0113B
!!!YEM: Copyright 1995, estate of Helmut Schaffrath.
**kern
*ICvox
*Ivox
*M4/4
*k[b-]
*F:
=1
{4cc
4cc
4a
4f
=2
2cc
4cc}
{4a
=3
4g
4f
4f
4e
=4
2f
2r}
==
!!!AGN: erzaehlendes Volks - Lied, Schalk -, Schelmen - Lied, betrogene Liebe, Ballade ?
!!!ONB: ESAC (Essen Associative Code) Database: ERK1
!!!AMT: simple quadruple
!!!AIN: vox
!!!EED: Helmut Schaffrath
!!!EEV: 1.0
*-
```

Figure 5: Humdrum Representation of the Folk Song ”Nun schürz dich, Gretlein”; Signature: deut0781

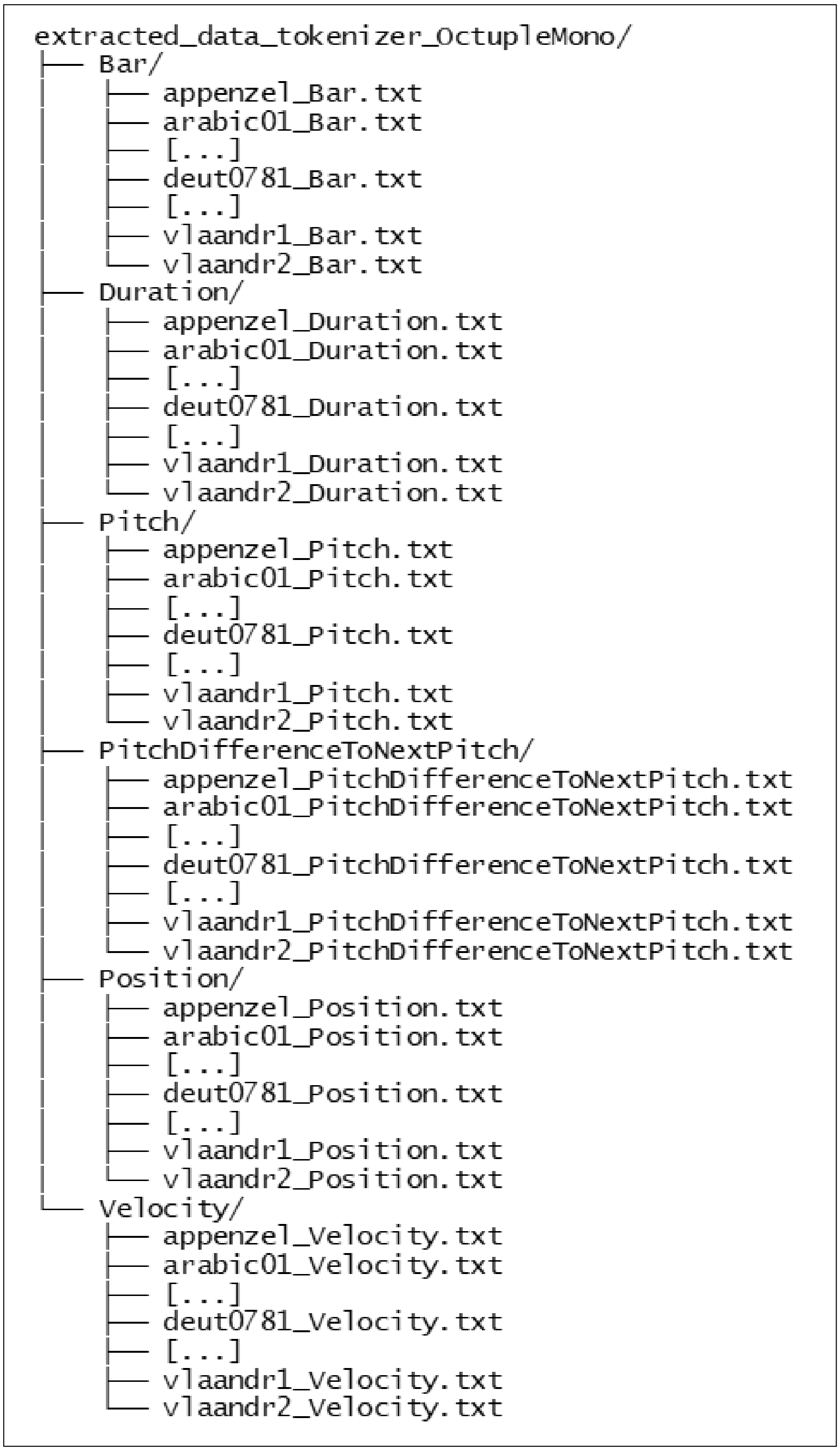


```
[TokSequence(tokens=[
'Pitch_72', 'Velocity_91', 'Duration_1.0.8', 'TimeShift_1.0.8',
'Pitch_72', 'Velocity_91', 'Duration_1.0.8', 'TimeShift_1.0.8',
'Pitch_69', 'Velocity_91', 'Duration_1.0.8', 'TimeShift_1.0.8',
'Pitch_65', 'Velocity_91', 'Duration_1.0.8', 'TimeShift_1.0.8',
'Pitch_72', 'Velocity_91', 'Duration_2.0.8', 'TimeShift_2.0.8',
'Pitch_72', 'Velocity_91', 'Duration_1.0.8', 'TimeShift_1.0.8',
'Pitch_69', 'Velocity_91', 'Duration_1.0.8', 'TimeShift_1.0.8',
'Pitch_67', 'Velocity_91', 'Duration_1.0.8', 'TimeShift_1.0.8',
'Pitch_65', 'Velocity_91', 'Duration_1.0.8', 'TimeShift_1.0.8',
'Pitch_65', 'Velocity_91', 'Duration_1.0.8', 'TimeShift_1.0.8',
'Pitch_64', 'Velocity_91', 'Duration_1.0.8', 'TimeShift_1.0.8',
'Pitch_65', 'Velocity_91', 'Duration_2.0.8'],
ids=[55, 114, 131, 196, 55, 114, 131, 196, 52, 114, 131, 196, 48, 114,
131, 196, 55, 114, 139, 204, 55, 114, 131, 196, 52, 114, 131, 196, 50,
114, 131, 196, 48, 114, 131, 196, 48, 114, 131, 196, 47, 114, 131, 196,
48, 114, 139],
bytes=None, events=[Event(type=Pitch, value=72, time=0, desc=72), […]],
ids_bpe_encoded=False, _ids_no_bpe=None)]
```

Figure 6: “Structured” Token Representation (One-Dimensional) of the Folk Song ”Nun schürz dich, Gretlein”; Signature: deut0781

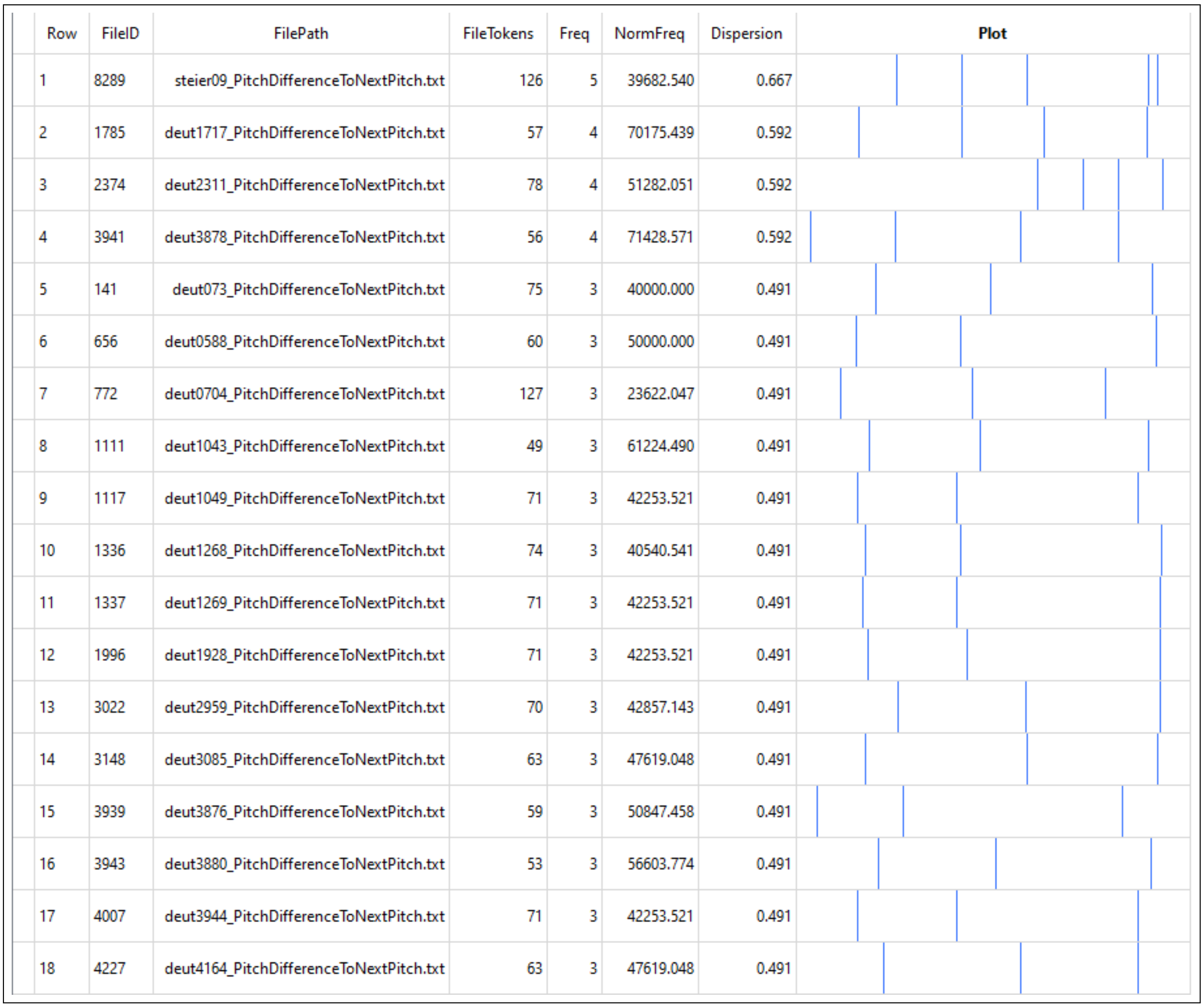
```
[TokSequence(tokens=
[['Pitch_72', 'Velocity_91', 'Duration_1.0.16', 'Position_0', 'Bar_0'],
['Pitch_72', 'Velocity_91', 'Duration_1.0.16', 'Position_16', 'Bar_0'],
['Pitch_69', 'Velocity_91', 'Duration_1.0.16', 'Position_32', 'Bar_0'],
['Pitch_65', 'Velocity_91', 'Duration_1.0.16', 'Position_48', 'Bar_0'],
['Pitch_72', 'Velocity_91', 'Duration_2.0.16', 'Position_0', 'Bar_1'],
['Pitch_72', 'Velocity_91', 'Duration_1.0.16', 'Position_32', 'Bar_1'],
['Pitch_69', 'Velocity_91', 'Duration_1.0.16', 'Position_48', 'Bar_1'],
['Pitch_67', 'Velocity_91', 'Duration_1.0.16', 'Position_0', 'Bar_2'],
['Pitch_65', 'Velocity_91', 'Duration_1.0.16', 'Position_16', 'Bar_2'],
['Pitch_65', 'Velocity_91', 'Duration_1.0.16', 'Position_32', 'Bar_2'],
['Pitch_64', 'Velocity_91', 'Duration_1.0.16', 'Position_48', 'Bar_2'],
['Pitch_65', 'Velocity_91', 'Duration_2.0.16', 'Position_0', 'Bar_3']],
ids=[[55, 26, 19, 4, 4], [55, 26, 19, 20, 4], [52, 26, 19, 36, 4], [48,
26, 19, 52, 4], [55, 26, 35, 4, 5], [55, 26, 19, 36, 5], [52, 26, 19, 52,
5], [50, 26, 19, 4, 6], [48, 26, 19, 20, 6], [48, 26, 19, 36, 6], [47, 26,
19, 52, 6], [48, 26, 35, 4, 7]],
bytes=None, events=None, ids_bpe_encoded=False, _ids_no_bpe=None)]
```

Figure 7: “OctupleMono” Token Representation (Two-Dimensional) of the Folk Song ”Nun schürz dich, Gretlein”; Signature: deut0781



```
extracted_data_tokenizer_OctupleMono/
├── Bar/
│   ├── appenzel_Bar.txt
│   ├── arabic01_Bar.txt
│   ├── [...]
│   ├── deut0781_Bar.txt
│   ├── [...]
│   ├── vlaandr1_Bar.txt
│   └── vlaandr2_Bar.txt
├── Duration/
│   ├── appenzel_Duration.txt
│   ├── arabic01_Duration.txt
│   ├── [...]
│   ├── deut0781_Duration.txt
│   ├── [...]
│   ├── vlaandr1_Duration.txt
│   └── vlaandr2_Duration.txt
├── Pitch/
│   ├── appenzel_Pitch.txt
│   ├── arabic01_Pitch.txt
│   ├── [...]
│   ├── deut0781_Pitch.txt
│   ├── [...]
│   ├── vlaandr1_Pitch.txt
│   └── vlaandr2_Pitch.txt
├── PitchDifferenceToNextPitch/
│   ├── appenzel_PitchDifferenceToNextPitch.txt
│   ├── arabic01_PitchDifferenceToNextPitch.txt
│   ├── [...]
│   ├── deut0781_PitchDifferenceToNextPitch.txt
│   ├── [...]
│   ├── vlaandr1_PitchDifferenceToNextPitch.txt
│   └── vlaandr2_PitchDifferenceToNextPitch.txt
├── Position/
│   ├── appenzel_Position.txt
│   ├── arabic01_Position.txt
│   ├── [...]
│   ├── deut0781_Position.txt
│   ├── [...]
│   ├── vlaandr1_Position.txt
│   └── vlaandr2_Position.txt
└── Velocity/
    ├── appenzel_Velocity.txt
    ├── arabic01_Velocity.txt
    ├── [...]
    ├── deut0781_Velocity.txt
    ├── [...]
    ├── vlaandr1_Velocity.txt
    └── vlaandr2_Velocity.txt
```

Figure 8: Organized Folder Structure of Extracted Data, Segregated by Parameter and Musical Piece



Row	FileID	FilePath	FileTokens	Freq	NormFreq	Dispersion	Plot
1	8289	steier09_PitchDifferenceToNextPitch.txt	126	5	39682.540	0.667	
2	1785	deut1717_PitchDifferenceToNextPitch.txt	57	4	70175.439	0.592	
3	2374	deut2311_PitchDifferenceToNextPitch.txt	78	4	51282.051	0.592	
4	3941	deut3878_PitchDifferenceToNextPitch.txt	56	4	71428.571	0.592	
5	141	deut073_PitchDifferenceToNextPitch.txt	75	3	40000.000	0.491	
6	656	deut0588_PitchDifferenceToNextPitch.txt	60	3	50000.000	0.491	
7	772	deut0704_PitchDifferenceToNextPitch.txt	127	3	23622.047	0.491	
8	1111	deut1043_PitchDifferenceToNextPitch.txt	49	3	61224.490	0.491	
9	1117	deut1049_PitchDifferenceToNextPitch.txt	71	3	42253.521	0.491	
10	1336	deut1268_PitchDifferenceToNextPitch.txt	74	3	40540.541	0.491	
11	1337	deut1269_PitchDifferenceToNextPitch.txt	71	3	42253.521	0.491	
12	1996	deut1928_PitchDifferenceToNextPitch.txt	71	3	42253.521	0.491	
13	3022	deut2959_PitchDifferenceToNextPitch.txt	70	3	42857.143	0.491	
14	3148	deut3085_PitchDifferenceToNextPitch.txt	63	3	47619.048	0.491	
15	3939	deut3876_PitchDifferenceToNextPitch.txt	59	3	50847.458	0.491	
16	3943	deut3880_PitchDifferenceToNextPitch.txt	53	3	56603.774	0.491	
17	4007	deut3944_PitchDifferenceToNextPitch.txt	71	3	42253.521	0.491	
18	4227	deut4164_PitchDifferenceToNextPitch.txt	63	3	47619.048	0.491	

Figure 9: Results (Excerpt) Created by Using the Plot-Tool (Search Term = -2.0 -2.0 -1.0 -2.0 -2.0) on the Custom Database

MITRA-zh: An efficient, open machine translation solution for Buddhist Chinese

Sebastian Nehrdich[1], **Marcus Bingenheimer**[2], **Justin Brody**[3], **and Kurt Keutzer**[1]

[1]University of California, Berkeley , Berkeley Artificial Intelligence Research (BAIR)
[2]Temple University, Philadelphia,
[3]Franklin and Marshall College, Lancaster

Abstract

Buddhist Classical Chinese is a challenging low-resource language that has not yet received much dedicated attention in NLP research. Standard commercial machine translation software performs poorly on this idiom. In order to address this gap, we present a novel dataset of 209,454 bitext pairs for the training and 2.300 manually curated and corrected bitext pairs for the evaluation of machine translation models. We finetune a number of encoder-decoder models on this dataset and compare their performance against commercial models. We show that our best fine-tuned model outperforms the currently available commercial solutions by a considerable margin while being much more cost-efficient and faster in deployment. This is especially important for digital humanities, where large amounts of data need to be processed efficiently for corpus-level operations such as topic modeling or semantic search. We also show that the commercial chat system GPT4 is surprisingly strong on this task, at times reaching comparable performance to our finetuned model and clearly outperforming standard machine translation providers. We provide a limited case study where we examine the performance of selected different machine translation models on a number of Buddhist Chinese passages in order to demonstrate what level of quality these models reach at the moment.

1 Introduction

Regarding the languages of the Buddhist tradition, there is a striking gap between the amount of available material in their ancient source languages Pāli, Sanskrit, Buddhist Classical Chinese, and Tibetan, and the number of available translations into Western languages, leaving the majority of texts inaccessible to a wider audience. In the case of Buddhist Chinese, only about 10% of the digitally available material has been translated into Western languages over the last two centuries. Machine translation (MT) therefore holds a great promise to help in the process of translating these texts, which is proceeding at a slow pace so far. Translators of Buddhist Chinese texts into Western languages generally do not work with MT tools yet, as their performance is rather poor. For common tasks in digital humanities such as topic modeling or semantic comparison, MT models also hold great promise as they make it possible to apply models trained primarily on English to this material. On the technical level, recent years have brought substantial advances in the performance of data-efficient MT systems, and their high level of reliability for language directions with good resources such as French or German to English has led to their wide adoption. In the case of low-resource languages with only limited data resources, the training of stable and usable MT systems remains difficult. This paper deals with the challenging situation of Buddhist Classical Chinese, which has been generally neglected in the compilation of openly available large parallel datasets such as OPUS[1] or the training of multilingual MT models such as NLLB (Team et al., 2022). We make the following contributions to this problem:

1. The first description of a dedicated Buddhist Classical Chinese to English parallel dataset with a total number of 209,454 bitext pairs covering a big variety of genres, including a manually curated and post-corrected evaluation dataset.

2. Two augmentation strategies using domain-specific feature engineering to

[1] https://opus.nlpl.eu/

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 266–277
December 1-3, 2023 ©2023 Association for Computational Linguistics

increase the performance of encoder-decoder models on this task.

3. Fine-tuning and evaluation of a variety of different openly available MT models as well as commercial providers on this dataset, giving the first thorough assessment of the quality of different currently available solutions on this language. We make the best-performing fine-tuned model available publicly.

4. Analysis of the behavior of these translation systems on different domains of Buddhist Chinese using standard MT metrics, followed by a more careful analysis using in-domain knowledge of Buddhist Classical Chinese.

In section 2 we give an overview of the relevant literature. Section 3 describes the datasets as well as the two augmentation strategies. In section 4 we show the evaluation results of the different models. In section 5 we examine the performance on different domains and conduct the case study.

1.1 Buddhist Chinese

For the purpose of this paper, we take ”Buddhist Chinese” to be the language of pre-modern Chinese Buddhist texts, both those translated from Indian originals and Buddhist texts composed directly in Chinese. Buddhist Chinese is a subset of Classical Chinese characterized morphologically by its frequent use of polysyllabic terms (usually translations of Indian words). Syntactically, translated texts in Buddhist Chinese often retain traces of Indian syntax and other grammatical features. Structurally, some of the texts mix prose and verse in a way that is characteristic for Indian literature, but highly unusual for non-Buddhist Classical Chinese (at least in the first millennium). In addition, compared to Classical Chinese in general, Buddhist Classical Chinese often preserves vernacular elements and has been used to study the development of Middle Chinese in the first millennium (Anderl, 2017). We estimate that there are between 8,000 and 10,000 extant Buddhist Chinese texts of which about 6000 have been digitized as full text. These texts were translated or authored by Chinese, Indian, Korean, and Japanese writers between the second and the nineteenth century. As the holy scriptures of Chinese, Korean, Japanese, and Vietnamese Buddhists, these texts are a highly significant part of the East Asian cultural heritage and are still widely used and studied.

Translation from Buddhist Chinese into various European languages began slowly in the 19th century. Currently, the largest bibliography (Bingenheimer, 2023) lists 1452 translations of 650 texts. I.e. over the course of some 200 years c. 10% of the digitally available published corpus has been translated into Western languages. The picture looks very different for Korean and Japanese. In Japanese, there is a translation of the Taishō canon in 355 vols. (Kokuyaku issaikyō 国訳一切経, 1930-1988) and a number of other large translation collections. A full translation into modern Korean exists of an influential 13th-century edition of the Buddhist canon (Dongguk yŏkkyong wŏn 東國譯經院, 1964-2001), which has been fully digitized[2].

2 Related Work

Along with natural language processing in general, the field of neural MT was revolutionized by the introduction of the transformer architecture in 2017 (Vaswani et al., 2017). These networks introduced a way of processing language that emphasized determining which parts of a sentence should be *attended to*. By processing a number of tokens simultaneously (rather than sequentially), transformers are capable of learning relations between more distant parts of a sentence than had been possible in widely used models like LSTMs. The resulting revolution in NLP is ongoing, with current iterations of models like OpenAI’s ChatGPT hardly needing any introduction.

Transformers come in 3 broad flavors: encoder-only, decoder-only, and encoder-decoder. Decoder-only models such as GPT are trained to predict the next token in a partially completed sequence, while encoder-decoder models learn to encode inputs and then decode them appropriately.

In this paper, we will use decoder-only and encoder-decoder variants, with mBART50 (Tang et al., 2021), WMT21 (Tran et al., 2021)

[2] https://kabc.dongguk.edu

and NLLB (Team et al., 2022) being encoder-decoder models while LlaMA2 is decoder-only.

The first of these, mBART50, utilizes monolingual pretraining with a denoising objective. NLLB on the other hand is trained directly on a massive multilingual parallel dataset without a pretraining stage. The WMT21 model we use consists of dense English to many and many to English models for translating between various languages (including Chinese). The open-source LlaMA2 (Touvron et al., 2023) is a decoder-only models which is pretrained on massive monolingual datasets consisting primarily of English data.

We decided to work with mBART and NLLB since both models have shown significant jumps in performance on low-resource languages when compared to randomly initialized transformer configurations. We will not use mT5 (Xue et al., 2021), which follows a similar pretraining objective as mBART, in our evaluation scheme since mBART has shown to be of equal or slightly better performance on low-resource translation tasks (Lee et al., 2022). The first and to our knowledge only dedicated publication on the problem of Buddhist Classical Chinese MT is (Li et al., 2022). Unfortunately, they have not made their models or evaluation data available, making it impossible to compare their findings in this paper.

3 Dataset

We collect a total number of 209,454 bitext pairs/5,738,025 characters from a variety of texts of the Taishō canon that have been translated into English. Due to the mechanics of the tokenizers of common language models, a single Chinese character is roughly equal to one token. The detailed genre distribution is given at table **??**. The composition dates of the Chinese texts in this dataset range from about ca. 150 to 1600 CE, while the English translations have been composed between ca. 1900 and 2020 CE. The training dataset covers a wide variety of different Buddhist Chinese domains: early and Mahāyāna sūtras, canon law, philosophical treatises, commentaries, ritual texts, etc. About half of the texts are translations of Indic Buddhist sources, others were composed directly in Buddhist Classical Chinese. Due to the diachronic spread and the variety of genres, the language of the Chinese Buddhist corpus is quite varied. Like there was no standard glossary to translate Indian terms into Chinese, English translations from Chinese were never standardized. Thus for any one Indic Buddhist term, we usually have a variety of different renderings in Chinese and English (Sanskrit āyatana, Buddhist Chinese: 處, 界, 入, English: sphere, field, sense organ, sense object, stage, level, base of cognition, sense sphere etc.). These circumstances do not only create challenges for the training of a MT system but also for the automatic evaluation of their performance, as in many cases, multiple different translations for a given Buddhist Chinese term are valid.

Category	Translated	Total	(%)
T01-0151 Āgama	516.570	2.861.382	18.1
T0152-0219 Past Lives	280.605	2.695.950	10.4
T0220-0261 Perfection of Wisdom	39.477	6.896.505	0.6
T0262-0277 Lotus Sūtra	133.955	587.509	22.8
T0278-0309 Flower Garland	1.052.629	2.262.554	46.5
T0310-0373 Treasure Trove	50.599	2.070.942	2.4
T0374-0396 Great Final Nirvāṇa	126.955	1.207.887	10.5
T0397-0424 Great Collection	37.109	1.566.096	2.4
T0425-0847 Sūtra	173.345	5.459.878	3.2
T0848-1420 Tantra	197.859	5.058.930	3.9
T1421-1504 Vinaya	122.236	5.264.857	2.3
T1505-1535 Sūtra Commentaries	352.091	2.046.519	17.2
T1536-1563 Abhidharma	503.450	5.125.379	9.8
T1564-1578 Madhyamika	26.720	417.713	6.4
T1579-1627 Yogācāra	331.903	2.506.840	13.2
T1628-1692 Collection of Treatises	257.920	1.202.910	21.4
T1693-1803 Chinese Commentaries	10.351	11.612.367	0.1
T1851-2025 Chinese Sectarian Writings	153.942	7.357.158	2.1
T2026-2120 History/Biography	599.823	6.235.952	9.6
T2121-2136 Encyclopedias/Dictionaries	1.350.386	2.255.979	59.9

Table 1: Distribution of the training data in the dataset according to different categories. "Translated" indicates the number of translated characters, "Total" indicates the total number of characters in the given Taishō section. The last column indicates how much of a section is available in translation. We only give Taishō sections that actually have translations into English.

As a first step, the English translations have been digitized and optical character recognition has been applied when necessary in order to obtain machine-readable unicode text. The translations have then been aligned with their Chinese counterparts as contained in the Chinese Buddhist Electronic Text Association (CBETA) corpus.[3] Many CBETA texts are based on an early 20th-century canonical edition, the ”Taishō Canon”. The texts have been thoroughly proofed against their original print editions. In some cases, punctuation has been added which increases

[3] http://cbeta.org/

intelligibility for humans (at least). Since on average the language models that we are evaluating have been exposed to more training data in simplified CJKV characters than in traditional CJKV, we convert the characters in the dataset to simplified characters during preprocessing.

The alignment was performed with vecalign (Thompson and Koehn, 2019). The embedding model used for the alignment process is a modified version of the multilingual sentence embedding model LaBSE (Feng et al., 2022). This model was further finetuned on a small corpus of gold-quality Buddhist Chinese to English bitext pairs. In order to reduce the influence of misaligned sentences, we use a rule-based scheme to remove sentences where the aligned English section is either much shorter or much longer than the Chinese counterpart. We also exclude samples from the training process where the Chinese part is shorter than four characters, assuming that NMT models do not learn well from very short samples.

We manually curated an evaluation dataset with a total size of 2,300 bitexts from a variety of texts from different genres of the Buddhist Chinese canon. Passages of the following texts have been included: T0026 (447 bitexts), T0374 (518 bitexts), T0475 (185 bitexts), T1585 (307 bitexts), T1600 (784 bitexts), T1970 (234 bitexts), T2062 (66 bitexts). The alignment of the evaluation dataset was performed by vecalign and then manually post-corrected. Training and evaluation data can be made available on request.

3.1 Data Augmentation

One commonly used strategy to improve MT performance with low-resource languages is the generation of synthetic training data that can be used to augment the original dataset during the training process. One synthetic augmentation technique for NMT systems is backtranslation (Sennrich et al., 2016), in which a large corpus in the target language is translated into the source language with the help of a MT system, creating a dataset where one side is automatically produced, and this data is then included in the training of the model. In our case, backtranslation has not proven to be helpful. While a lot of English data in the target language exists that could potentially be used, this data is not in the desired target domain, and utilizing this data results in significantly lower performance.

We therefore propose two different strategies for augmentation of the Buddhist Classical Chinese dataset:

1. Creation of a synthetic Classical Chinese to English dataset by machine-translating the Modern Chinese sentences of the NiuTrans Classical Chinese to Modern Chinese dataset.[4] we use the multi-lingual Transformer model of the Meta-AI research group submitted to the WMT2021 shared task, wmt21-dense-24-wide-x-en with 4.7billion parameters (Tran et al., 2021), to translate the Modern Chinese into English. We decided to use this model as among the openly available translation models, this has shown the best performance for Modern Chinese to English translation for this domain. This generates a total number of 972,470 bitext pairs. We make this dataset available at https://github.com/dharmamitra/NiuTrans-Classical-Modern-English.

2. Prompting ChatGPT3.5 with Buddhist Chinese paragraphs together with their translation into Modern Korean and dictionary entries in order to create a synthetic Buddhist Chinese to English dataset.

ChatGPT3.5 is a large language model created by OpenAI. Its most recent, more expensive version is GPT4. We utilize the digitally available complete translation of the Chinese Buddhist canon into Korean[5] in order to train a mBART (Lewis et al., 2020) Buddhist Chinese to Korean translation model. Then, to create additional data, we take random pseudo-paragraphs of up to 200 characters in length together with their Korean translation obtained via the mBART model and feed them to ChatGPT3.5, prompting it to translate the pseudo-paragraph into English, making use of the Korean translation. We also augment the prompt with dictionary entries obtained via the Digital Dictionary of Buddhism[6] (Muller, 2019) to ensure better translation of specific Bud-

[4] https://github.com/NiuTrans/Classical-Modern

[5] https://kabc.dongguk.edu/

[6] http://www.buddhism-dict.net/

dhist terminology. In order to avoid overgeneration of possible entries, we limit the retrieval to entries that are three characters or longer. We prompt ChatGPT to output the English translation together with the Chinese source sentences in a sentence-aligned format, thus generating as many sentence pairs as are needed to meet our desired augmentation target. In this way, we generate a total number of 436,945 synthetic Buddhist Chinese to English sentence pairs. We make this dataset available at https://github.com/dharmamitra/buddhist-chinese-agumentation.

4 Experiments

We evaluate the following models: Bing Translator[7], DeepL[8] and Google Translate[9] are commercial translation engines. We test ChatGPT3.5 and GPT4 are the commercial chat systems provided by OpenAI. We query the OpenAI models with a simple prompt: ”*Translate the following Buddhist Chinese passage into English: <sentence> English:*”. Transformer 600M serves as the baseline for the finetuned models, which is the NLLB600M model with randomly initialized weights, simulating the training of a transformer model with 600M parameters from scratch. mBART50 and mBART50-to-1 are two different versions of the multilingual BART model with a size of 611M parameters (Tang et al., 2021). Both are pretrained on a denoising task, while the latter is the many–to-one version that is finetuned on a many-to-one translation task, including Chinese, with English as the target language. No further information is provided during the prompting step. NLLB600M-3.3B is the massive multilingual model of Meta AI in different sizes, trained among other languages also on Chinese to English. WMT21 is the Meta AI's submission to WMT21 News Translation task (Tran et al., 2021). We use the wmt21-dense-24-wide-x-en version with 4.7B parameters, which was also trained on the Chinese to English task.

We fine-tuned the 7B parameter version of LlaMA2 available on HuggingFace, using QLoRA for fintetuning on the full parallel dataset. During training and inference, we used the prompt *"Below is some text in Classical Chinese. It is taken from the Buddhist literature. Write a translation of the text into English."* followed by labelled Chinese inputs and a label for the English translation.

[7] https://www.bing.com/translator

[8] https://www.deepl.com/translator

[9] https://translate.google.com/

Model	BLEU	chrF++
Bing Translator sent	4.1	25.5
Bing Translator par	4.4	27.9
Deepl sent	7.6	30.1
DeepL par	8.2	33.0
Google Translate sent	8.5	31.8
Google Translate par	8.9	35.1
ChatGPT3.5 sent	9.5	35.0
ChatGPT3.5 par	11.2	38.8
GPT4 sent	11.8	37.4
GPT4 par	12.8	40.3
Transformer 600M sent	4.4	31.0
Transformer 600M par	7.7	35.8
mBART50 sent-ft	11.2	35.4
mBART50 par-ft	11.6	37.5
mBART50-to-1 sent	3.3	22.3
mBART50-to-1 sent-ft	13.0	37.8
mBART50-to-1 par-ft	12.9	39.4
NLLB600M sent	2.0	19.0
NLLB600M sent-ft	12.6	37.0
NLLB600M par-ft	13.3	39.6
NLLB1.3B sent-ft	13.5	38.4
NLLB1.3B par-ft	13.8	40.0
NLLB3B sent-ft	14.6	39.5
NLLB3B par-ft	14.4	40.9
WMT21 sent	5.1	26.1
WMT21 sent-ft	14.2	38.7
WMT21 par-ft	14.4	41.0
WMT21+aug sent-ft	**15.2**	**39.9**
WMT21+aug par-ft	**15.1**	**41.7**
LlaMA2-ft sent	8.6	31.8
LlaMA2-ft par	8.8	32.8

Table 2: Main results on the MT task. Models finetuned with our parallel data are indicated with ft. Sent indicates evaluation on sentence-level, par indicates evaluation on paragraph-level.

We finetune all encoder-decoder models on sentence-level and on pseudo-paragraph-level as we assume that a larger context might help the models to arrive at better translation solutions. For the pseudeo-paragraph level, we concatenate adjecent sentences with a total length of up to 200 tokens. We decided on this number as the encoder-decoder model with the shortest context length, WMT21, only supports up to 200 tokens. We did a thorough hyperparameter search on a fixed holdout set to determine the optimal learning rate and number of training steps for each model.

4.1 Evaluation

We present the results in table 2. We evaluate using two different metrics: BLEU (Papineni et al., 2002) which uses word-level n-grams and chrF++ (Popović, 2017), which works with character-level n-grams. Since English translations of Buddhist Classical Chinese works frequently use borrowed terms from Sanskrit where different writing conventions might be applied, chrF++ seems a more appropriate choice as it considers similarity on character-level, and not just on word-level as is the case with BLEU. We do not use model-based metrics such as COMET or BERTscore as they have not been finetuned on the Buddhist domain and we can therefore not assume that they are appropriate for this scenario.

Regarding the commercial providers Bing, DeepL and Google Translate, their results are clearly inferior to those of ChatGPT3.5 and GPT4. The weak score of Bing Translate is especially remarkable in light of the fact that it was marketed to explicitly support Literary Chinese.[10] GPT4 in turn performs better than ChatGPT3.5 with a clear margin. All commercial systems perform better when the data is provided on pseudo-paragraph level instead of sentence level. The baseline model Transformer 600M struggles to reach usable performance. Also the openly available models that have been trained on the Chinese-to-English MT objective, mBART50-to-1, NLLB, and WMT21, perform badly without finetuning, being clearly inferior even when compared to DeepL and Google Translate. After finetuning on our dataset, they show a significant performance boost and clearly outperform the baseline Transformer 600M, showing that denoising pretraining in the case of mBART50 and transfer learning from Modern Chinese to English in cases of the other models is beneficial for this specific task. This is also confirmed by the fact that finetuned mBART50-to-1 performs better than finetuned mBART50 by 2.4 chrF++ score on sentence and 1.9 on paragraph level, which further proves that a model pretrained on the denoising objective benefits from further finetuning on the Modern Chinese to English translation task before being finetuned on our dataset. The fact that the NLLB models all perform bet-

[10] https://www.microsoft.com/en-us/translator/blog/2021/08/25/microsoft-translator-releases-literary-chinese-translation/

ter than mBART50 further supports the observation that transfer learning from Modern CHinese to English is helping. In the NLLB family, we see a clear improvement of performance with increasing model size. Noteworthy is the fact that while the smallest model NLLB600M benefits significantly from pseudo-paragraph-level training with an increase in BLEU of 0.7 and in CHRF of 2.6, the performance of the 3B version is not better in terms of BLEU, while better in terms of CHRF with an increase of 1.4. It is therefore safe to conclude that the increase of model performance by pseudo-paragraph level training decreases with model size for the NLLB family. The largest model that we finetune, WMT21 with 4.7B parameters, shows the best zero-shot performance of all openly available models. The finetuned version of this model shows almost identical performance with NLLB3B-ft on pseudo-paragraph level while being slightly inferior on sentence level. When we add the augmentation data to this model, we see a visible improvement of 1.2 chrF++ score on sentence level and 0.7 chrF score on the pseudo-paragraph level, leading to the highest performance of all models evaluated in this paper. Since the training with the augmentation data is very resource- and time-consuming, we could not evaluate its effects on the behavior of the other models. LlaMA2 does not yet competitive performance after finetuning on the dataset as it performs poorer than the finetuned mBART50 andNLLB600M models.

Text	Model	BLEU	chrF++
T0026	ChatGPT3.5	12.0	39.1
	GPT4	13.9	40.8
	WMT21+aug ft	**17.8**	**43.4**
T0374	ChatGPT3.5	12.0	39.0
	GPT4	13.6	40.4
	WMT21+aug ft	**15.6**	**42.4**
T0475	ChatGPT3.5	12.8	39.6
	GPT4	**13.6**	**41.0**
	WMT21+aug ft	11.9	38.0
T1585	ChatGPT3.5	9.5	38.1
	GPT4	12.2	40.7
	WMT21+aug ft	**19.9**	**48.9**
T1600	ChatGPT3.5	10.9	40.0
	GPT4	**12.3**	**41.1**
	WMT21+aug ft	12.2	39.8
T1970	ChatGPT3.5	9.6	37.9
	GPT4	11.4	39.1
	WMT21+aug ft	**12.0**	**40.3**
T2026	ChatGPT3.5	9.6	38.0
	GPT4	**11.6**	**39.4**
	WMT21+aug ft	11.0	37.9

Table 3: Performance on individual texts of the evaluation dataset. All results are calculated on pseudo-paragraph level.

5 Analysis

Table 3 shows the performance of ChatGPT3.5, GPT4, and WMT21+aug-ft on different evaluation texts measured in BLEU and chrF++ on pseudo-paragraph level. WMT21+aug-ft outperforms the other models on T0026, T0374, T1585, and T1970. For T0026 and T1585, the difference to the second best-performing model GPT4 is significant. In the case of those texts where GPT4 performs better than WMT21+aug-ft, the difference is generally small, with T0475 being the only exception. ChatGPT3.5 performs worse than GPT4 on all texts, and, again with the exception of T0475, ChatGPT3.5 also performs worse than WMT21+aug-ft on all texts. The reason for GPTx output being comparatively strong on T0475-par might be because the text, the Vimalakīrtinirdeśa, is available online in a number of different versions, while most of the other texts in the evaluation set have only been translated from Chinese to English only once so far. It is known that GPTx is trained on large amounts of online data and therefore, memorization of the evaluation data is a possibility here. Compared to T0475, translations of the other texts are rather more recent and not as readily available online.

In order to understand the nature of the mistakes that the different translation models produce, we analyzed several passages manually. We give the full samples in the appendix. The first paragraph is taken from T1585, the Cheng weishi lun, a core text of Sino-Indian Yogācāra philosophy. On this text, WMT21+aug-ft shows a generally superior quality, producing less serious mistakes, which mirrors the BLEU score results on the individual texts. It is noteworthy that while all three models on average use the right vocabulary to translate the philosophical terms in this paragraph, ChatGPT3.5

struggles significantly and GPT4 struggles somewhat to interpret the dense syntax of the Chinese. WMT21+aug-ft is doing visibly better, but certain points of confusion remain, i.e. rendering 非無 (here: ”not nonexistent”) as ”neither nonexistent nor existent”, which is not correct.

For T0026, an early Buddhist sūtra text, the BLEU and chrF++ score does not well align with our manual evaluation. Although the metric indicates a clear advantage for WMT21+aug-ft, many passages are actually rendered more accurately in the GPT4 output. It is possible that the metric was influenced by the tendency of WMT21+aug-ft to use Sanskrit terms for their Chinese equivalents, something that is common practice in Buddhist translation. In our example, the five great rivers of India (Jambudvīpa) are mentioned and while the GPTx models render 恒伽，搖尤那，舍牢浮，阿夷羅婆提，摩企 with at times misleading pinyin transcriptions (Hengqie/Hengqia, Shalao Fu/Sheloufu etc.), WMT has ”Ganges, the Yamunā, the Śrāvastī, the Ajiravatī, and the Mahī”. Śrāvastī is a mistake for Sarabhū here, but one can see how the Sanskrit terms in the output might influence the n-gram based BLEU and chrF++ scores, which compare it to the human reference translation that has similar terms.

T2062 is an early 17th-century biography of a Chinese monk. Next to T1970 (a 12th century Pure Land treatise) it is the text in our sample for which the linguistic markers of ”Buddhist Chinese” are least evident. It is thus not that surprising that our domain-specific model does not produce significant differences to GPT4 for those two texts. The language of T2062 differs from that of the other evaluation texts in that it contains many named entities, esp. person and place name, which are often referenced in an abbreviated way. The syntax is exceedingly terse with almost no redundancy or repetition. Overall all models performed worst on T2062, with many passages translated wrongly to a degree that post-editing means retranslation. The example in the appendix is atypical in that it compares a relatively ”easy” passage, which all models have managed to render reasonably well. As we have seen with T0026, WMT21+aug-ft tries to identify Sanskrit terms and render them as such (Jambudvīpa), but in this case unsuccessfully (ch. 荼毗, skr. kṣapita). An interesting passage that shows how the context understanding of ChatGPT3.5 is inferior to GPT4 and WMT21+aug-ft is 所聞種種，隨力不同 (“[all people] smelled something different, according to their powers [of insight]”). Although in itself its choices are reasonable, ChatGPT3.5 misses the subject with ”Various sounds and scents were heard, depending on the strength of the fire.” Note how the ambivalence of 聞 throws the model off. 聞 can indeed mean ”to hear” or ”to smell”, but not both at the same time, in English ”sounds and scents were heard” is nonsense.

Compared with the two large proprietary GPTx models, the domain-specific WMT21+aug-ft model shows at least approximately equal, and often better, BLEU and chrF++ scores. It needs to be remembered that inference on commercial GPTx models costs orders of magnitude more than the fine-tuned WMT21 model, which can be efficiently served even on a single consumer-grade GPU (Peng et al., 2023). Another problem when interacting with commercial models is the fact that their performance has been shown to differ significantly even in a relatively short amount of time, making their behavior unpredictable (Chen et al., 2023).

6 Conclusion

For this paper we compiled a novel dataset for the training and evaluation of MT models for Buddhist Classical Chinese. We applied two methods of data augmentation and compared a number of different encoder-decoder models finetuned on this data against large commercial MT providers and commercial decoder-only chat models. The domain-specific evaluation as well as the BLEU/chrF++ scores show that with the help of the augmentation strategies, our much smaller and locally run model WMT21+aug-ft clearly outperforms the standard commercial providers as well as the commercial chat system ChatGPT3.5, while being on par with GPT4 and outperforming it on certain domains. Significantly, in the case of texts in the Chinese canon that are originally translated from Indic sources, our finetuned model outperforms GPT4 and is therefore the currently best available solution. For Chinese-Chinese Buddhist texts, the performance of GPT4 is comparable to our models. This makes the finetuned model an ideal solution for semantic similarity tasks on corpus level, which are of central concern within the digital humanities and require cost-efficient and fast processing of large quantities of data.

The evaluation results show that WMT21+aug-ft as well as GPT4 have reached a level of maturity that for the first time in the history of the translation of Buddhist Chinese texts into Western languages, these tools can be of genuine help to translators.

We see a number of directions for further work:

First, the amount of digitized English translations aligned with their Buddhist Chinese counterparts is still very limited. Increasing the size of this dataset promises further improvements in performance especially when it comes to Chinese-Chinese material, for which there is less bitext in the current dataset.

Second, while we present two strategies for data augmentation in this paper that are clearly boosting the performance of encoder-decoder models, further refinements of these approaches, especially the prompting of commercial engines with the right prior data, promise further significant leaps in performance.

Third, while LlaMA2 has not shown competitive performance in our evaluation, the preliminary results are encouraging. By utilizing more monolingual data during the finetuning stage, we might see significant performance increases for smaller, open decoder-only models as well. Fourth, our limited manual examination of the output of the MT models has indicated that widely used evaluation methods such as BLEU or chrF++ do not always align well with human judgment, an observation that was made in other recent studies with a focus on large language models on MT as well (Wang et al., 2023). We therefore see a clear need for a thorough examination of alternative evaluation methods in future studies.

References

Christoph Anderl. 2017. Medieval Chinese Syntax, volume 2, pages 689–703. Brill.

Marcus Bingenheimer. 2023. Bibliography of translations (by human translators) from the chinese buddhist canon into western languages.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is chatgpt's behavior changing over time?

Dongguk yŏkkyong wŏn 東國譯經院. 1964-2001. Hangul daejang kyong.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

En-Shiun Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Adelani, Ruisi Su, and Arya McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? In Findings of the Association for Computational Linguistics: ACL 2022, pages 58–67, Dublin, Ireland. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.

Denghao Li, Yuqiao Zeng, Jianzong Wang, Lingwei Kong, Zhangcheng Huang, Ning Cheng, Xiaoyang Qu, and Jing Xiao. 2022. Blur the linguistic boundary: Interpreting chinese buddhist sutra in english via neural machine translation. In 2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI), pages 228–232.

A. Charles Muller. 2019. The Digital Dictionary of Buddhism and CJKV-English Dictionary: A Brief History. In Digital Humanities and Buddhism: An Introduction, SISAP 2013, A Coruña, Spain, October 2-4, 2013, Proceedings, pages 143–156.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation.

Maja Popović. 2017. chrF++: words helping character n-grams. In Proceedings of the Second Conference on Machine Translation, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3450–3466, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved Sentence Alignment in Linear Time and Space. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook ai ’ s wmt21 news translation task submission. In Proc. of WMT.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. arXiv preprint arXiv:2304.02210.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer.

A Appendices

A.1 Case Study Full Examples

Double underline: serious mistakes. Single underline: less serious mistakes.

T1585: 復有迷謬唯識理者，或執外境如識非無、或執內識如境非有、或執諸識用別體同、或執離心無別心所

Human Reference: Some are deluded about this principle of consciousness only. Some grasp the external world as not nonexistent, like consciousness. Some grasp internal consciousness as nonexistent, like the external world. Some grasp the various kinds of consciousness as being the same in substance but different in characteristics. Some believe that there are no separate mental functions apart from the mind.

ChatGPT3.5: There are those who hold mistaken views regarding the theory of the only consciousness. Some believe that external objects are not without consciousness, while others believe that internal consciousness is not the same as external objects. Some believe that various consciousnesses have different natures, while others believe that there is no separate mind apart from the mind.

GPT4: Furthermore, there are those who are confused and mistaken about the principles of consciousness-only, some clinging to the idea that external objects exist as consciousness and are not non-existent, some clinging to the idea that internal consciousness exists as objects and are not non-existent, some clinging to the idea that all consciousnesses function with the same substance, and some clinging to the idea that there is no separate mental object apart from the mind.

WMT21+aug-ft: Again, there are those who are deluded and mistaken about the principle of consciousness only. They may hold that external objects, like consciousness, are neither nonexistent nor existent, or they may hold that internal consciousness, like objects, is neither existent nor existent, or they may hold that the functions of the various forms of consciousness are different but the substance is the same, or they may hold that apart from mind there are no separate mental activities.

T0026: 是時，晃昱諸天始生天者，不諳世間成敗，不見世間成敗，不知世間成敗，見大火已，皆恐怖毛竪而作是念：『火不來至此耶？火不來至此耶？』前生諸天諳世間成敗，見世間成敗，知世間成敗，見大火已，慰勞諸天曰：『莫得恐怖，火法齊彼，終不至此。』

Human Reference: Then the gods of radiance, who have just been born as gods and have never experienced, seen, or known the world’ s [cyclic] creation and destruction, on seeing that great fire will all be terrified, with their hair standing on end, and will think: ”The fire will not reach up to here, will it? The fire will not reach up to here, will it?” Those who were born as gods at an earlier time and have experienced, seen, and known the world’ s [cyclic] creation and destruction will, on seeing that great fire, calm the distress of those gods, telling them, ”Don’ t be afraid! That fire is of such a nature that it will not reach up to here.”

GPT-3.5: At that time, the heavenly beings who were born in the beginning of the world were ignorant of the success or failure of the world. They did not see the success or failure of the world, nor did they know the success or failure of the world. When they saw the great fire, they were all terrified and had the following thought: ”Has the fire reached here? Has the fire reached here?” The heavenly beings from previous lives who were knowledgeable about the success or failure of the world, who saw the success or failure of the world, and who knew the success or failure of the world, comforted the other heavenly beings, saying: ”Do not be afraid. The fire will not reach here. It will be extinguished before it reaches here.”

GPT-4: At that time, the newly born gods, who were not familiar with the success and failure of the world, who did not see the success and failure of the world, and who did not know the success and failure of the world, saw the great fire and were all terrified, their hairs standing on end, and thought: ”Will the fire come here? Will the fire come here?” The gods who had been born earlier, who were familiar with the success and failure of the world, who saw the success and failure of the world, and who knew the success and failure of the world, saw the great fire and comforted the other gods, saying: ”Do not be afraid, the law of fire is equal to that, it will never come here.”

WMT21+aug-ft: At that time, the Ābhāsvara gods who had just been reborn among the gods did not study the formation and decay of the world, did not see the formation and decay of the world, and did not know the formation and decay of the world. But when they saw the great fire, they were all terrified, their hair standing on end, and they thought, ‘Will the fire not come here? Will the fire not come here? ’ The gods who had been reborn previously studied the formation and decay of the world, saw the formation and decay of the world, and knew the formation and decay of the world. But when they saw the great fire, they comforted the gods, saying, ‘Do not be afraid! The nature of fire is equal to that. It will never come here.’

T2062 Zhenqing: 荼毗日天色霽明，淨無纖翳，舉火之際，忽有片雲，如蓋凝覆其上，灑微雨數點。烟焰起時，異香充塞，內自殿閣僧房，外自路人船子，所聞種種，隨力不同。火餘骨有三色，而鏘鏘有聲。

Human reference translation: On the day of his cremation, the weather had cleared up, and not a cloud was in the sky. But the moment the fire was lighted, a little cloud appeared suddenly. It grew denser just above the cremation site, like a canopy, and gently sprinkled a few drops of rain. When the fire arose, an extraordinary scent filled the air, and everybody, those in the monastery halls and monks’ quarters, as well as those outside, the people on the roads and in their boats, all smelled something different, according to their powers [of insight]. The bones remaining in the ashes were of three colors and made a clear sound [when tapped].

ChatGPT3.5: On the day of cremation, the sky cleared up, pure and without any blemish. Just as the fire was being lit, suddenly a patch of clouds appeared, covering and sprinkling a few drops of light rain. As the smoke and flames rose, a strange fragrance filled the air, coming from the inner halls and monk’s quarters as well as from passersby and boats on the road. Various sounds and scents were heard, depending on the strength of the fire. After the fire subsided, the bones were left with three different colors and made a clanging sound.

GPT4: On the day of the cremation, the sky was clear and bright, without a trace of cloud. At the moment the fire was lit, a patch of cloud suddenly appeared, covering the area like a canopy, sprinkling a few drops of light rain. When the smoke and flames rose, a strange fragrance filled the air, permeating from the inner temple halls and monks’ quarters to the outer passersby and boats. Various reactions were heard, differing according to one’s perception. The remaining bones after the fire had three colors, and they made a clinking sound.

WMT21+aug-ft: The day of Jambudvīpa The sky was clear and clean, without the slightest cloudiness. While the fire was being raised, a cloud suddenly appeared, like a canopy, covering the fire and sprinkling a few drops of rain. When the smoke and flames arose, they were filled with an extraordinary fragrance. From inside the palace and the monks’ quarters, and from outside the boats of the passers-by, the fragrance varied according to one’ s ability to smell it. The embers left by the fire were of three colors, and they emitted a sound when burned.

Comparison on Heterosexual and Homosexual Woman's Lonely Heart Ads in Taiwan: Taking *AllTogether* and *Lesbian* Board on PTT Web Forum as Examples

Yu-Hsuan Lin
National Taiwan University / No. 1, Sec. 4, Roosevelt Road, Taipei, 10617 Taiwan
linsarahyh@gmail.com

Abstract

This study aims to compare the lonely heart ads of heterosexual and homosexual women in Taiwan. The data was collected from the *AllTogether* (heterosexual) and *Lesbian* (homosexual) boards on the PTT web forum. Word frequency analysis and topic modeling were used to analyze the data. It was found that lesbians tend to state more on emotional and spiritual connection, using words that describe personality traits and changes in emotions. Heterosexual women, on the other hand, showed more concern about practical matters such as religion, occupation, and habits, possibly with the goal of building a family relationship through marriage and starting a family.

1 Introduction

Dating and mating are important research areas in many disciplines, including sociology and gender studies. Numerous studies discuss heterosexual mating preferences from various perspectives, comparing the differences between heterosexual men and women. In recent years, there have also been more and more research focus on homosexual partner choice. Among these studies, we can also see some studies comparing mating preferences between different sexual orientations, although the number is limited, it is not a completely unexplored area, particularly in the West. For example, Potârcă et al. (2015) showed that social tolerance and legal recognition of same-sex marriage are positively correlated with higher intentions of long-term dating and a stronger belief in monogamy.

However, when we focus on Taiwan, the first country in Asia to legalize same-sex marriage, there are very few such studies. Shieh and Tseng (2010) note that previous research on partnership formation and choice in Taiwan has mainly focused on the description of heterosexual relationships, and has been mostly interpreted through biological evolutionism, social learning theory, and social exchange theory to explain the concept of "marry within one's social class"(門當戶對) as well as marriage gradient. So far, there are less than ten papers in Taiwan that focus on homosexual mating preference, and there is almost no literature that compares the differences in partner choice among different sexual orientations. Therefore, in an aim to bridge this research gap I will compare more than 2,000 lonely heart ads of heterosexual and homosexual women published in AllTogether and Lesbian discussion boards of the PTT web forum from 2021 to 2022 and will reflect on the original heterosexual-based mating preference research and gender assumptions in Taiwan and how it compares to similar homosexual data, specifically lesbians.

2 Background

Taiwan has been a trailblazer for LGBTQ+ rights in Asia, legalizing same-sex marriage in 2019. Many researchers attribute the relatively positive attitude towards the LGBTQ+ community in Taiwan to how *queer theory* was one of the key theories flourishing in Taiwan's academic and cultural spheres in the 1990s and helped transform the public discourse on sexuality and gender in media and popular culture early on (Liou, 2005; Guo, 2023). However, just 15 years ago, it was still difficult to recruit LGBTQ+ participants for studies due to fear of stigmatization even in Taiwan (Shieh, 2010). More recent studies discuss the Lesbian communities in Taiwan more intimately by examining the outward expression of masculine inclinations, specifically *zhongxing* – gender neutrality in appearance –, and its relation to normative constraints and queer agency (Hu, 2019).

In Taiwanese lesbian culture, labels such as T, P, or versatile are frequently used. T stands for "tomboy," typically referring to individuals who exhibit relatively masculine traits in personality and appearance (e.g., very short hair). P comes from the

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 278–287
December 1-3, 2023 ©2023 Association for Computational Linguistics

Chinese word “老婆(lǎopó)”, which means wife or female spouse, signifying a more feminine quality compared to T. Versatile, in Chinese word "不分", refers to a state between T and P. Hu (2018)’s research highlighted how these labels dominate the realms of self-identity and community interaction among Taiwanese lesbians. With the influence of Western queer theory and deconstructionism entering Taiwan in the 1990s, these labels have also been influenced in terms of their usage and definitions. Traditionally, T and P are often seen as a pair of lesbian roles, but various combinations, such as T-T, P-P, and diverse self-labeling (e.g., "masculine P", "long-haired T", and "versatile-P"), have become common expressions in the community.

There are several papers that focus on comparing women of different sexualities’ choice of partners (see e.g. Potârcă et al. 2015 and Ybarra and Mitchell, 2016). Russock (2011) analyzed personal advertisements with an evolutionary interpretation and found that heterosexual women had a tendency of resource-seeking and offered more physical attractiveness. A study by Veloso et al. (2014) also showed that heterosexual women emphasize the characteristics of good provision of resources and emotional investment in long-term relationship. Also, they point out that homosexual women’s partner selection pattern resembles to both heterosexual man and woman. But Willis (2014) stated that lesbians have higher expectations for their female mates than straight men have for theirs, and women are preferred over men for emotional and social needs regardless of sexuality. Shieh and Tseng (2010) interviewed 33 homosexual couples in Taiwan in order to find their mate choosing preference, found that except for their homosexual identity and shared socialization experiences, their preferences are almost identical to those of heterosexual counterparts. Li and Lu (2020)’s analysis focused on the Desiring Self and Desiring Others texts of Taiwanese lesbians and gay men on dating websites. He pointed out that even in same-sex relationships, there continues to be an expectation of pairs of masculine/feminine and active/passive roles. This implies that homosexual communities still have mate expectations influenced by traditional gender roles associated with norms in heterosexual relationships. However, overall, there is relatively limited research both domestically and internationally that specifically explores and compares the dating preferences and differences among women with different sexual orientations and the underlying reasons for these preferences.

3 Data and Method

3.1 Data Selection

PTT Bulletin Board System is the largest web forum in Taiwan. According to the Institute for Information Industry (for Information Industry, 2017), PTT is one of the most used social websites in Taiwan after Facebook, Instagram, and YouTube. The main language used on PTT is traditional Chinese. There are many boards in PTT, and different interests and topics are gathered under each board. *Alltogether* is a board about matchmaking, which consists mainly of lonely heart ads primarily written by heterosexual men and women who are looking for a partner. The *Lesbian* board is a lesbian-oriented community that includes renting together, chat group recruitment, news, and lonely heart ads. These lonely heart ads usually consist of a long self-introduction and some ideal criteria for a future partner. PTT is also easy to crawl for its text-based interface and does not limit crawling or scraping.

The paper employs a Python script with ’BeautifulSoup’ library to scrape data from PTT. The script iterates through *Alltogether* and *Lesbian*, extracts data from each article, including author names, board names, article titles, publication times, and content. The data used in this study consists of all lonely heart ads, whose titles included “徵男“ (looking for males) on Alltogether, and “自介” (self-introduction) and “她介” (introducing my friend) on Lesbian, posted between 2021-2022. A total of 899 posts were collected from Lesbian and 1472 posts were collected from Alltogether.

3.2 Segmentation, POS Tagging & Stop word removal

Segmentation and POS tagging were conducted through CKIP (Yang and Lin, 2023), a natural language processing tool for traditional Chinese language developed by Academia Sinica in Taiwan. A list of stop words was filtered out to make the analysis process more efficient.

3.3 Counting Frequency

The basis of this study is a simple statistical comparison of the word frequency in both sets of data. First, the number of occurrences of each word in the text was calculated and then normalized by dividing the total number of occurrences of the word by the total number of words to get the percentage of the word in the text. Next, I compared the words that appeared more than 10 times on both boards. If the word is more than 2 times more likely to appear in one of the boards than the other, it is judged to be the more used word of that specific board.

3.4 Topic Modeling

Topic modeling is a statistical model used to discover abstract topics in a series of documents. It assumes that each document is associated with one or more topics and that each topic will have a corresponding word distribution. I used a simple Latent Dirichlet Allocation (LDA) model to process the data for this study. Only words whose POS tagged by CJIP are non-predicate adjective (A), common noun (Na), proper noun (Nb), place noun (Nc), nominalized verb (Nv), active intransitive verb (VA), active causative verb (VAC), active pseudo-transitive verb (VB), active transitive verb (VC), stative intransitive verb (VH), stative pseudo-transitive verb (VI), and stative transitive verb (VJ) were included. I used topic modelling to try to identify the differences in topics that homosexual and heterosexual women talk about when seeking a spouse.

4 Results

The calculation of word frequency shows that there are indeed many differences in word usage between lesbian and heterosexual females. The more used words of lesbian woman can be divided into several categories including but not limited to the following:

1. Nouns related to sexual orientation. For example, “女朋友”(girlfriend), “性別”(sexuality), and “伴侶”(partner).

2. Nouns referring to emotion and thoughts, such as “共鳴” (resonance), “情緒” (emotion), “理想” (ideal), “靈魂” (soul), and “心事” (have something on one’s mind).

3. Nouns related to the length of hair.

4. Explicit and clear adjectives to describe personal characteristics and relationships.

5. Verbs to describe inner feelings, especially about love. For instance, “吸引” (attract), “愛上” (fell in love with), “嚮往” (long for), and “樂意” (willing to).

6. Verbs describing future prospects, such as “走過” (walk through), “尋找” (look for), “開” (start), “邁向”(toward), and “建立” (build).

For heterosexual women, their more used words can be divided into several categories including but not limited to the following:

1. Nouns related to males.

2. Nouns of region name.

3. Nouns of real-life information, such as “宗教” (religion), “寵物” (pet), “房貸” (housing loan), and “職業” (job).

4. Verbs describe living habits and practical future plans of forming a family, such as “抽菸” (smoking), “煮飯” (cook), and “生子” (give birth).

5. A series of descriptive adjectives that do not involve actual details. From “強” (better; strong), “良好” (well), “正常” (normal), to “不良” (bad).

6. Adjectives about body image, which are “胖” (fat) and “瘦” (thin) mostly.

4.1 Inconclusive results from topic models

As for topic modeling, it is unfortunate that the calculated keyword combinations, despite achieving a coherence score of 0.48 in text from Lesbian and 0.4 from Alltogether, are difficult to discern the correlation and consistency between them. These texts could not be successfully classified into several clear themes thus we can not even make a comparison between the two sets. The possible reason is that lonely heart ads themselves are already a very specific category. They consisted mainly of self-introduction and requirements for the future partner. According to the previous frequency analysis, though, we can find that the words and concerns of women with different sexual orientations do differ, the LDA algorithm itself has a larger granularity and is perhaps more suited for classifying articles compared to a subdivision of content within the same topic.

5 Conclusion and Discussion

According to the result of counting word frequency, we find that lesbians are more concerned with spiritual, emotional, and inner communication and connection. They have more adjectives that explicitly describe their (or their ideal partner's) personalities, as well as verbs and nouns that describe change of heart states. They have more spiritual interaction requirements for their potential partners. Lesbians use many verbs that describe abstract nouns to illustrate the stages and progress of the relationship. In contrast, heterosexual women care about more practical issues, they reveal themselves or ask potential partners for more details about their place of residence, occupation, religion, and so on. There is also more information and requirements about having a specific habit or not, for example, drinking and smoking. Also, marriage and giving birth are also more used words by heterosexual women, all of these imply that heterosexual women's lonely heart ads are not only about finding a partner for love but also about building a family relationship based on a permanent life together. However, they use many adjectives that do not have a clear statement of the object, such as well, good and bad, to state something about themselves or their counterparts, as compared to homosexual women's use of adjectives.

In addition, there are differences in the vocabulary related to physical appearance in the texts of the two sexual orientations. Terms like hair, long hair, and short hair are frequently mentioned in lesbian texts. Li and Lu (2020)'s research also identified this phenomenon, noting that the term "long hair" often appeared positively in the narratives on Taiwanese lesbian dating webstes between 2013 and 2015. It was frequently mentioned alongside other feminine attributes and was highly favored. This suggests that long hair and the femininity it symbolizes were preferred qualities. Seven years later, the data on PTT forum still found that hair length remained an important label used by lesbians for dating and self-introduction. Hair length and the gender qualities it represents continue to be significant factors in lesbian dating preferences. On the other hand, terms like "fat" or "thin" are frequent in vocabulary related to appearance in heterosexual women's texts. This indicates that body shape and its association with attractiveness may be more important in heterosexual relationships.

The findings on heterosexual women seem to be consistent with past findings that women are more concerned with resource exchange in relationships. However, the comparison with lesbians allows us to question the way in which this result has been interpreted in terms of biological evolutionary theory in the past. If this condition is particularly pronounced for heterosexual women only, is it a consequence of the social construction of heterosexual romantic relationships?

Lastly, the main limitation of this study are the sampling of data and the limitations of textual analysis itself. For the former, although PTT is the largest online forum in Taiwan, there are still many other dating apps and physical dating events. The people gathered on PTT must have a certain degree of bias compared to the population as a whole, rather than a random sampling. As an online community, PTT is likely to have its tendency of age group, political stance, and so on. In addition, as a single ad contains both self-introduction and requirements for potential mates, we can only make a tentative distinction in the analysis process by inference due to the lack of a fixed format for the postings. As for the latter, textual analysis's limitation lies in the diversity of norms and expectations among different groups, influencing what individuals choose to disclose or withhold in their written texts. While lesbians in the study appeared to prioritize spiritual aspects over material considerations compared to heterosexual women, it does not imply lower material preferences for lesbians or vice versa. These differences may stem from different communication cultures, etc. It's crucial not to equate textual expressions with actual mate selection criteria. To obtain a more comprehensive understanding, future research should integrate textual analysis with fieldwork and interviews, enabling a nuanced exploration of partner preferences within cultural and social contexts.

6 Future work

In order to have a better understanding of the mate preferences of Taiwanese of different sexual orientations, I think we can

1. cross-compare more commonly used words by including information of men

2. collect data from the dating app, which will give more categories on personal interests and expectations, and make it easier to carry out detailed topic model

3. interpret the study of mate preference into a larger context of "doing gender" for discussion and comparison.

7 Code

Please find all the code at the following url: https://github.com/hiitslin/taiwan_les_and_hetero_comparison

References

Institute for Information Industry. 2017. Bacheng yishang taiwanren aiyong facebook. line zuowen shequnwangzhan longtou. yiren pingjun yong shige shequn zhanghao. nian qingren gengai youtube he ig[more than 80% of taiwanese love facebook. line is the leading social networking site. one person has 4 social media accounts on average. young people prefer youtube and ig]. Institute for Information Industry.

Wangtaolue Guo. 2023. Queers on the move: sinicizing queer theory and theorizing queerness in taiwan. *Perspectives*, 31(2):220–234.

Y.-Y Hu. 2018. Compelling categories, shifting identities: Social media, transnational cultural politics, and "t/po/bufen" lesbian identity formation. *Taiwan Journal of Anthropology*, 16:1–50.

Yu-Ying Hu. 2019. Mainstreaming female masculinity, signifying lesbian visibility: The rise of the zhongxing phenomenon in transnational taiwan. *Sexualities*, 22(1-2):182–202.

Po-Wei Li and Chia-Rung Lu. 2020. Articulating sexuality, desire, and identity: A keyword analysis of heteronormativity in taiwanese gay and lesbian dating websites. *Sexuality and Culture*, 24:1499–1521.

Liang-ya Liou. 2005. Queer theory and politics in taiwan: the cultural translation and (re) production of queerness in and beyond taiwan lesbian/gay/queer activism. *NTU studies in Language and Literature*, 14:123–154.

Gina Potârcă, Melinda Mills, and Wiebke Neberich. 2015. Relationship preferences among gay and lesbian online daters: Individual and contextual influences. *Journal of Marriage and Family*, 77(2):523–541.

H. I. Russock. 2011. An evolutionary interpretation of the effect of gender and sexual orientation on human mate selection preferences, as indicated by an analysis of personal advertisements. *Behaviour*, 148(3):307–323.

W Shieh and H. Tseng. 2010. Entering into couple relationship: An explorative study of gay and lesbian partner-selection preference in taiwan. *The Journal of Guidance Counseling*, 32(2):27–46.

Wen-Yi Shieh. 2010. Gay and lesbian couple relationship commitment in taiwan: A preliminary study. *Journal of Homosexuality*, 57(10):1334–1354.

Vivianni Veloso et al. 2014. Comparison of partner choice between lesbians and heterosexual women. *Psychology*, 5(02):134.

Jarryd T Willis. 2014. Partner preferences across sexual orientations and biological sex. *Personal Relationships*, 21(1):150–167.

Mu Yang and Li-Huai Lin. 2023. Ckip transformers for traditional chinese.

Michele L Ybarra and Kimberly J Mitchell. 2016. A national study of lesbian, gay, bisexual (lgb), and non-lgb youth sexual behavior online and in-person. *Archives of sexual behavior*, 45(6):1357–1372.

A Appendix

A.1 More Commonly Used Words for the *Lesbian* Category

Original Word	**English**
伴侶	partner
女孩	girl
女朋友	girlfriend
姊姊	older sister
少女	girl
性	sex
性別	sexuality
老婆	wife
議題	issue (often refers to social aspect)

Table 1: Nouns related to sexual orientation

Original Word	**English**
傷害	harm
內心	in one's heart
共鳴	resonance
困擾	harass
心事	something on one's mind
心思	mood; thought
情感	emotion
感受	feeling
煩惱	worries

Table 2: Nouns refers to emotion and thoughts

Original Word	English
髮	hair
短髮	short hair
長髮	long hair

Table 3: Nouns related to length of hair

A.2 Part of Classified More Used Words on Alltogether

Original Word	English
(VH)上進	enterprising
(VH)中性	neutral
(VH)可靠	dependable
(VH)合得來	hit it off
(VH)多元	diverse
(VH)孤獨	lonely
(VH)安心	at ease; relieved
(VH)專注	concentrated
(VH)帥氣	good-looking (refers to male)
(VH)平淡	dull; ordinary (refers to lifestyle)
(VH)廣泛	wide range of
(VH)悶騷	mild on the outside but wild on the inside
(VH)正經	serious (refers to personality)
(VH)浪漫	romantic
(VH)清秀	decent looking
(VH)溫柔	gentle
(VH)無趣	boring
(VH)獨特	unique
(VH)相似	similar
(VH)真實	real; true
(VH)純	pure
(VH)細水長流	small but steady stream (refers to condition of relationship)
(VH)細膩	attentive; considerate
(VH)舒適	comfortable
(VH)親密	intimate; close
(VH)誠懇	sincere
(VH)變好	getting better
(VH)貼心	thoughtful
(VH)驚喜	surprise

Table 4: Explicit and clear adjectives to describe personal characteristics and relationships

Original Word	English
佩	-like
勝過	better than
受	(passive voice expressions)
吸引	attract
失去	lose
愛上	fall in love with
沈迷	be addicted to
認同	identify with; agree to
同意	agree
喜歡上	like
嚮往	long for
感受	feel
感受到	feel
明白	understand
關心	care

Table 5: Verbs to describe inner feelings, especially about love

Original Word	English
尋找	look for
帶來	bring
建立	build
接近	get close to
擁抱	embrace
放下	let go
準備好	be prepared to
開啓	open; start
走過	walk through
邁入	toward
邁向	toward
交給	leave something to
探索	discover

Table 6: Verbs describe prospection to future

Original Word	English
男	male
男人	male
男生	boy
異性	opposite sex

Table 7: Nouns related to males

Original Word	English
宜蘭	Yilan
彰化	Changhua
新北	New Taipei
新竹	Hsinchu

Table 8: Nouns of name of region

Original Word	English
宗教	religion
寵物	pet
小孩	kid
年次	year of born
房貸	housing loan
業務	job description
碩士	master degree
習慣	habit
職業	job
股票	stock
規劃	plan (refers to future)
身高	height
公司	company
大學	university
學校	school

Table 9: Nouns related to real-life information

Original Word	English
抽煙	smoking
抽菸	smoking
減肥	lose weight
煮飯	cook
生子	give birth
結婚	getting married

Table 10: Verbs describe living habit and practical future plan of forming family

Original Word	English
不良	bad
佳	good
强	better; strong
有限	limited
正常	normal
正當	decent; legitimate
特殊	special
相對	relatively; comparatively
良好	well

Table 11: A series of descriptive adjectives that do not involve actual details

Original Word	English
胖	fat
瘦	thin

Table 12: Adjectives about body image

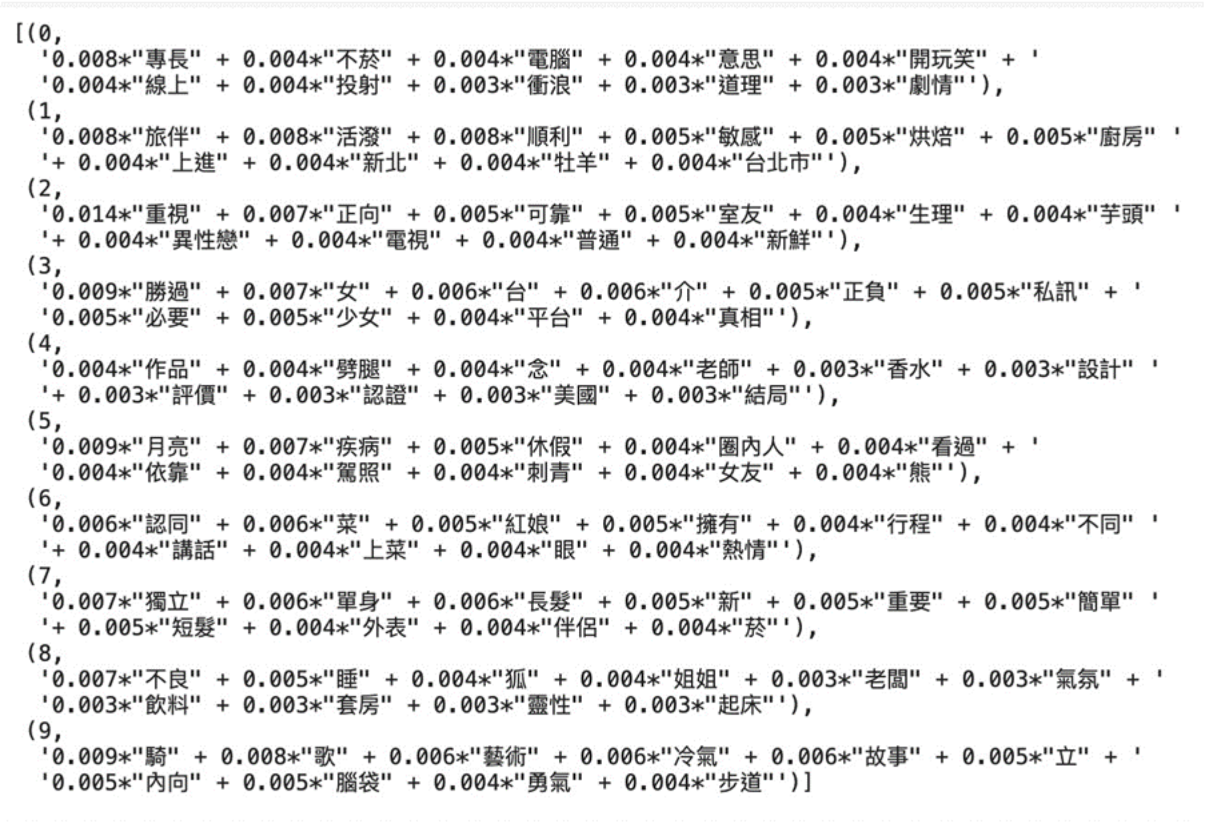
A.3 Result of LDA Model

Perplexity	-8.964043551
Coherence	0.4853863106

Table 13: Coherence score of LDA model of Lesbian

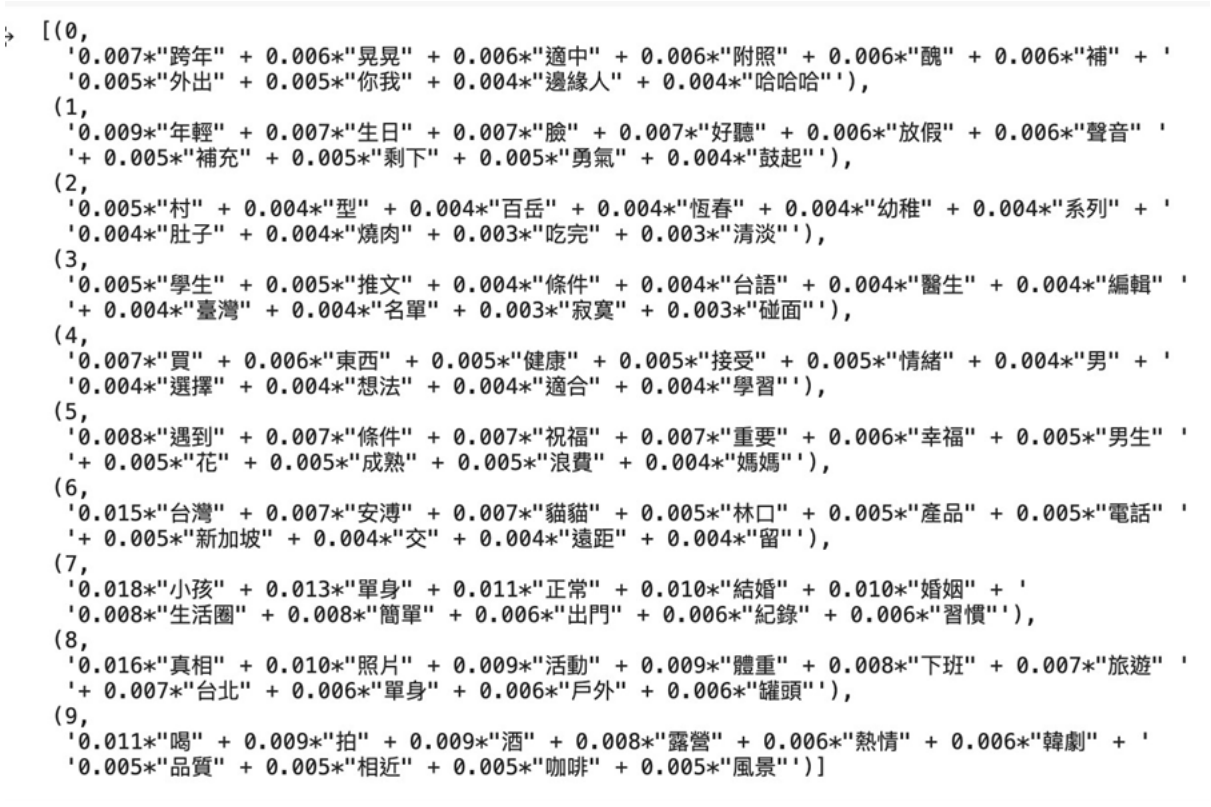
Perplexity	-8.96394623
Coherence	0.40502346209

Table 14: Coherence score of LDA model of Alltogether



```
[(0,
  '0.008*"專長" + 0.004*"不菸" + 0.004*"電腦" + 0.004*"意思" + 0.004*"開玩笑" + '
  '0.004*"線上" + 0.004*"投射" + 0.003*"衝浪" + 0.003*"道理" + 0.003*"劇情"'),
 (1,
  '0.008*"旅伴" + 0.008*"活潑" + 0.008*"順利" + 0.005*"敏感" + 0.005*"烘焙" + 0.005*"廚房" '
  '+ 0.004*"上進" + 0.004*"新北" + 0.004*"牡羊" + 0.004*"台北市"'),
 (2,
  '0.014*"重視" + 0.007*"正向" + 0.005*"可靠" + 0.005*"室友" + 0.004*"生理" + 0.004*"芋頭" '
  '+ 0.004*"異性戀" + 0.004*"電視" + 0.004*"普通" + 0.004*"新鮮"'),
 (3,
  '0.009*"勝過" + 0.007*"女" + 0.006*"台" + 0.006*"介" + 0.005*"正負" + 0.005*"私訊" + '
  '0.005*"必要" + 0.005*"少女" + 0.004*"平台" + 0.004*"真相"'),
 (4,
  '0.004*"作品" + 0.004*"劈腿" + 0.004*"念" + 0.004*"老師" + 0.003*"香水" + 0.003*"設計" '
  '+ 0.003*"評價" + 0.003*"認證" + 0.003*"美國" + 0.003*"結局"'),
 (5,
  '0.009*"月亮" + 0.007*"疾病" + 0.005*"休假" + 0.004*"圈內人" + 0.004*"看過" + '
  '0.004*"依靠" + 0.004*"駕照" + 0.004*"刺青" + 0.004*"女友" + 0.004*"熊"'),
 (6,
  '0.006*"認同" + 0.006*"菜" + 0.005*"紅娘" + 0.005*"擁有" + 0.004*"行程" + 0.004*"不同" '
  '+ 0.004*"講話" + 0.004*"上菜" + 0.004*"眼" + 0.004*"熱情"'),
 (7,
  '0.007*"獨立" + 0.006*"單身" + 0.006*"長髮" + 0.005*"新" + 0.005*"重要" + 0.005*"簡單" '
  '+ 0.005*"短髮" + 0.004*"外表" + 0.004*"伴侶" + 0.004*"菸"'),
 (8,
  '0.007*"不良" + 0.005*"睡" + 0.004*"狐" + 0.004*"姐姐" + 0.003*"老闆" + 0.003*"氣氛" + '
  '0.003*"飲料" + 0.003*"套房" + 0.003*"靈性" + 0.003*"起床"'),
 (9,
  '0.009*"騎" + 0.008*"歌" + 0.006*"藝術" + 0.006*"冷氣" + 0.006*"故事" + 0.005*"立" + '
  '0.005*"內向" + 0.005*"腦袋" + 0.004*"勇氣" + 0.004*"步道"')]
```

Figure 1: Result of LDA Model



```
[(0,
  '0.007*"跨年" + 0.006*"晃晃" + 0.006*"適中" + 0.006*"附照" + 0.006*"醜" + 0.006*"補" + '
  '0.005*"外出" + 0.005*"你我" + 0.004*"邊緣人" + 0.004*"哈哈哈"'),
 (1,
  '0.009*"年輕" + 0.007*"生日" + 0.007*"臉" + 0.007*"好聽" + 0.006*"放假" + 0.006*"聲音" '
  '+ 0.005*"補充" + 0.005*"剩下" + 0.005*"勇氣" + 0.004*"鼓起"'),
 (2,
  '0.005*"村" + 0.004*"型" + 0.004*"百岳" + 0.004*"恆春" + 0.004*"幼稚" + 0.004*"系列" + '
  '0.004*"肚子" + 0.004*"燒肉" + 0.003*"吃完" + 0.003*"清淡"'),
 (3,
  '0.005*"學生" + 0.005*"推文" + 0.004*"條件" + 0.004*"台語" + 0.004*"醫生" + 0.004*"編輯" '
  '+ 0.004*"臺灣" + 0.004*"名單" + 0.003*"寂寞" + 0.003*"碰面"'),
 (4,
  '0.007*"買" + 0.006*"東西" + 0.005*"健康" + 0.005*"接受" + 0.005*"情緒" + 0.004*"男" + '
  '0.004*"選擇" + 0.004*"想法" + 0.004*"適合" + 0.004*"學習"'),
 (5,
  '0.008*"遇到" + 0.007*"條件" + 0.007*"祝福" + 0.007*"重要" + 0.006*"幸福" + 0.005*"男生" '
  '+ 0.005*"花" + 0.005*"成熟" + 0.005*"浪費" + 0.004*"媽媽"'),
 (6,
  '0.015*"台灣" + 0.007*"安溥" + 0.007*"貓貓" + 0.005*"林口" + 0.005*"產品" + 0.005*"電話" '
  '+ 0.005*"新加坡" + 0.004*"交" + 0.004*"遠距" + 0.004*"留"'),
 (7,
  '0.018*"小孩" + 0.013*"單身" + 0.011*"正常" + 0.010*"結婚" + 0.010*"婚姻" + '
  '0.008*"生活圈" + 0.008*"簡單" + 0.006*"出門" + 0.006*"紀錄" + 0.006*"習慣"'),
 (8,
  '0.016*"真相" + 0.010*"照片" + 0.009*"活動" + 0.009*"體重" + 0.008*"下班" + 0.007*"旅遊" '
  '+ 0.007*"台北" + 0.006*"單身" + 0.006*"戶外" + 0.006*"罐頭"'),
 (9,
  '0.011*"喝" + 0.009*"拍" + 0.009*"酒" + 0.008*"露營" + 0.006*"熱情" + 0.006*"韓劇" + '
  '0.005*"品質" + 0.005*"相近" + 0.005*"咖啡" + 0.005*"風景"')]
```

Figure 2: LDA model of Alltogether

Author Index