Exploring Similarity Measures and Intertextuality in Vedic Sanskrit Literature

So Miyagawa
The University of Tsukuba
miyagawa.so.kb@u.tsukuba.ac.jp

Yuki Kyogoku
Leipzig University
kyogoku11@gmail.com

Yuzuki Tsukagoshi
The University of Tokyo
yuzuki@l.u-tokyo.ac.jp

Kyoko Amano
Kyoto University
amano.skskrt@kcn.jp

Abstract

This paper examines semantic similarity and intertextuality in selected texts from the Vedic Sanskrit corpus, specifically the Maitrāyaṇī Saṃhitā (MS) and Kāṭhaka Saṃhitā (KS). Three computational methods are employed: Word2Vec for word embeddings, stylo package for stylometric analysis, and TRACER for text reuse detection. By comparing various sections of the texts at different granularities, patterns of similarity and structural alignment are uncovered, providing insights into textual relationships and chronology. Word embeddings capture semantic similarities, while stylometric analysis reveals clusters and components that differentiate the texts. TRACER identifies parallel passages, indicating probable instances of text reuse. The computational analysis corroborates previous philological studies, suggesting a shared period of composition between MS.1.9 and MS.1.7. This research highlights the potential of computational methods in studying ancient Sanskrit literature, complementing traditional approaches. The agreement among the methods strengthens the validity of the findings, and the visualizations offer a nuanced understanding of textual connections. The study demonstrates that smaller chunk sizes are more effective for detecting intertextual parallels, showcasing the power of these techniques in unraveling the complexities of ancient texts.

1 Introduction

Vedic Sanskrit literature preserves invaluable cultural and historical information from ancient India. However, their study presents unique challenges due to linguistic characteristics, modes of composition, and transmission. Computational methods offer promising avenues to analyze such texts on an unprecedented scale. In this paper, we explore similarity measures and intertextuality between selected texts from the Vedic corpus - the Maitrāyaṇī Saṃhitā (MS) and Kāṭhaka Saṃhitā (KS). These texts belong to different śākhās or schools, and are considered to exhibit mutual influence in their composition around 900–700 BCE.[1]

The main focus of this paper is to present reliable numerical data on the chapter-wise similarity between the MS and KS. While it is known that the MS and KS are parallel texts, the variations in similarity among chapters have not yet been confirmed through numerical data. Since differences in chapter-wise similarity can contribute to estimating the relative chronology of each chapter, this similarity analysis holds significant importance for understanding the process of textual composition.

In recent years, the editorial process has been increasingly elucidated through philological studies (Amano, 2014-2015, 2020), suggesting variations in similarity between different sections in MS and KS depending on the time period. That is to say, sections edited in earlier periods exhibit lower similarity between MS and KS, whereas those edited in later times show higher similarity between MS and KS. Similarity analysis using computational methods further advances this study.

Our analysis employs three approaches:

1. Word embeddings generated using Word2Vec
2. Stylometry analysis using the stylo package
3. text reuse detection with TRACER

The word embeddings approach vectorizes the texts and compares the cosine similarity of the vectors. The stylo (Eder et al., 2016) and TRACER (Büchler, 2013; Büchler et al., 2018) approaches examine stylistic similarity and text reuse at document level.

The texts are pre-processed by undoing phonological change (sandhi) in the original texts and

[1] All the corpora, codes, and results are available on our GitHub repo https://github.com/somiyagawa/VedicSanskrit (accessed October 5, 2024).

lemmatizing the words. Different chunking of the text is compared — at section level and by segments of 20, 100 and 200 words.

The results demonstrate interesting patterns of similarity and clustering between different text segments, with general alignment between the three approaches. This research highlights the potential of computational methods in studying ancient languages and aims to inspire further collaborative research at the intersection of Indology and computational linguistics.

2 Related Work

Computational methods have been increasingly applied to study various aspects of Sanskrit literature in recent years. Hellwig et al. (2020) developed a neural network architecture for processing Sanskrit texts. Krishna et al. (2019) analyzed poetic style in Sanskrit poetry using deep learning techniques.

Regarding Vedic Sanskrit specifically, Hellwig et al. (2023) developed a dependency parser for R̥gvedic Sanskrit. Hellwig and Nehrdich (2018) compiled a Vedic treebank. These works provide NLP tools and resources for computational processing of Vedic texts.

Stylometry has been widely used to study authorship and stylistic similarity in classical literature. For instance, Stover et al. (2016) applied stylometric analysis using the stylo R package (Eder et al., 2016) to investigate the authenticity of an unknown classical Latin text called the *Expositio*. Their study concluded that this work was probably written by the second-century African author Apuleius of Madauros.

While stylometry focuses on authorship attribution and stylistic analysis on a macro-level, as demonstrated by the stylo package, text reuse detection tools offer a micro-level approach to detecting each text reuse such as quotations and allusions among texts. Specifically, TRACER (Büchler, 2013; Büchler et al., 2014; Büchler et al., 2018) is a text reuse detection tool that has been successfully applied to study intertextuality in ancient Greek (Buechler et al., 2008; Büchler et al., 2010), Latin (Franzini et al., 2018b), Coptic texts (Miyagawa, 2022, 2021; Miyagawa et al., 2018), Classical Tibetan (Almogi et al., 2019), German (Franzini et al., 2018a), etc.

Other programs are also available for historical text reuse analysis. For example, Tesserae and Passim are well-known tools in this field. Tesserae (Coffee et al., 2012) is primarily used for Latin texts, while Passim (Romanello and Hengchen, 2021) has been adapted for Western languages and Arabic with promising results but has not yet been adapted for Sanskrit. Compared to these tools, TRACER offers greater flexibility and customizability, making it possible to adapt it to Vedic Sanskrit using custom lemmatization, synonym, and cohyponym files.

3 Methodology

3.1 Corpus

The corpora consist of selected texts from the Maitrāyaṇī Saṃhitā (MS) and Kāṭhaka Saṃhitā (KS). The following sections are analyzed:

1. MS.1.1 (MS.1.1.1-1.1.13): 1145 words
2. MS.1.6 (MS.1.6.3-13): 3816 words
3. MS.1.7 (MS.1.7.2-5): 819 words
4. MS.1.9 (MS.1.9.3-8): 1627 words
5. KS.8 (KS.7.15 + 8.1-12): 3519 words
6. KS.9.1 (8.15 + 9.1-3): 818 words
7. KS.9.11 (9.11-17): 1721 words

The corpus MS.1.1 includes ritual formulas for new and full moon sacrifice. MS.1.6 includes ritual explanation about establishment of sacred fires, whose parallel is KS.8. MS.1.7 includes ritual explanation about reestablishment of sacred fires, whose parallel is KS.9.1. MS.1.9 includes explanation of secret spells related to ritualistic communal life, whose parallel is KS.9.11. MS.1.6, 1.7, and 1.9 have been philologically studied by Amano (2009), and their parallels in KS have been accurately identified. MS and KS were composed in the same editorial policy, and have almost the same contents for the same rituals, but with some variants in details. MS and KS contain portions with different linguistic styles and content (sometimes irregularly inserted or arranged), which necessitates the exclusion of such portions to conduct linguistically and semantically accurate analyses. The corpora used in this analysis were created to ensure that different styles (formulas or explanations) and contents (rituals) are not mixed. The chapter numbers within the parentheses following each section name represent the exact chapter numbers included in the section. The size (word count) of each corpus is also provided above.

Using these corpora, we conduct comparisons (similarity analyses) between sections as follows:

1. MS.1.1 ↔ MS.1.6
2. MS.1.6 ↔ MS.1.7
3. MS.1.6 ↔ KS.8
4. MS.1.7 ↔ KS.9.1
5. MS.1.9 ↔ KS.9.11

The first two comparisons, namely MS.1.1 ↔ MS.1.6 and MS.1.6 ↔ MS.1.7, serve as an evaluation of the proposed methods, as their similar or dissimilar relations are philologically demonstrated. MS.1.1 differs significantly in content from MS.1.6 and MS.1.7, while the latter two share similar contents. Accordingly, if the proposed methods work well, the comparison of MS.1.1 ↔ MS.1.6 is expected to show a low similarity, whereas MS.1.6 ↔ MS.1.7 is expected to demonstrate a high similarity, compared to the former comparison.

The following three comparisons are between the texts of MS and KS. Each chapter is thought to have been edited in different periods and under different cultural influence, and therefore, the degree of similarity between MS and KS varies. The two comparisons, MS.1.6 ↔ KS.8 and MS.1.7 ↔ KS.9.1, were manually calculated (Amano, 2014-2015). As a result, the comparison MS.1.6 ↔ KS.8 showed a low similarity, while the comparison MS.1.7 ↔ KS.9.1 showed a very high similarity. Since it is philologically inferred that MS.1.6 is older than MS.1.7, the paper presented the perspective that chapters compiled in the earlier period have lower similarity with KS, whereas those from a later period have higher similarity with KS, indicating possible intertextual borrowing. From this, in our current analysis, the comparison of MS.1.6 ↔ KS.8 is anticipated to reveal a low similarity, while MS.1.7 ↔ KS.9.1 is expected to exhibit a high similarity.

In contrast, the last comparison, MS.1.9 ↔ KS.9.11, was not examined in the previous studies, and serves as the main focus of our current analysis, aiming to demonstrate to what extent this comparison shows similarity. If the comparison of MS.1.9 ↔ KS.9.11 reveals a high similarity, akin to MS.1.7 ↔ KS.9.1, it strongly suggests that the intertextual contact between MS.1.9 and KS.9.11 occurred during a later period, characterized by a tendency for MS and KS to exhibit similarities, as argued in Amano (2020).

The texts are procured from the Digital Corpus of Sanskrit.[2] Original Sanskrit text undergoes phonetic fusion and changes at word boundaries, known as sandhi. These fusions and changes make it challenging to segment the text into individual words and perform morphological analysis. Therefore, as a first step in processing the text, it is necessary to resolve the sandhi to create an "un-sandhi-ed" text, which can then be used for lemmatization. The texts stored in the Digital Corpus of Sanskrit are processed into un-sandhi-ed texts as well as lemmatized texts by the computational method of (Hellwig et al., 2020) , verified through expert review and correction. For the purpose of comparing similarity, the lemmatized texts are used, which are manually divided into distinct chunks or paragraphs with attention to meaningful coherence:

- Section level
- Fixed-size segments of 20, 100 and 200 lemmas

3.2 Word Embeddings

Word embedding models capture semantic relationships between words from their co-occurrence in a large corpus. We use Word2Vec (Mikolov et al., 2013), a two-layer neural network that predicts surrounding context words given an input word. We employ the skip-gram library with the training algorithm set to the skip-gram and default parameters for other settings.

The training data consists of a collection of Vedic Sanskrit texts, excluding the MS and KS. The word embeddings are averaged for each segment to obtain a document vector. The similarity between document vectors is computed using cosine similarity. Cosine similarity is used to compare the document vectors by calculating the cosine of the angle between them. This measures how close the vectors are to each other while disregarding their magnitude.

3.3 Stylometry and Text Reuse

The stylometry analysis is performed using the stylo package in R (Eder et al., 2016). It supports a variety of statistical analyses to examine

[2] http://www.sanskrit-linguistics.org/dcs/index.php (accessed May 24, 2024)

stylistic similarity between texts, such as cluster analysis, multidimensional scaling, principal component analysis etc. We use the cosine similarity as the similarity metric. For text reuse detection, we use TRACER (Büchler, 2013; Büchler et al., 2014; Büchler et al., 2018), which has been successfully applied to study intertextuality in various ancient language corpora. It provides a Java implementation to detect different types of text reuse such as quotations, allusions and idioms.

4 Results

4.1 Word Embedding

Table 1 shows the average cosine similarity between text segments using Word2Vec. In general, the similarity scores increase as the chunk size increases from 20 to 200 lemmas.

Table 1: Average cosine similarity using Word2Vec

Text Pair	Chunk Size		
	20	100	200
MS.1.1 ↔ MS.1.6	0.813	0.899	0.925
MS.1.6 ↔ MS.1.7	0.856	0.934	0.959
MS.1.6 ↔ KS.8	0.863	0.941	0.964
MS.1.7 ↔ KS.9.1	0.860	0.940	0.971
MS.1.9 ↔ KS.9.11	0.844	0.933	0.959

The comparisons of MS.1.1 ↔ MS.1.6 exhibits a lower similarity than that of MS.1.6 ↔ MS.1.7, suggesting effective performance of the analysis. However, the high similarity of MS.1.6 ↔ KS.8, which were expected to less similar, contradicts the previous findings. This discrepancy from the expectation may arise from the larger number of dissimilar chunks compared to similar ones, despite the existence of parallels between MS.1.6 and KS.8. The dissimilarity is highlighted by averaging the similarity values, because even within sections that are considered to have high similarity, segments that do not correspond exhibit low similarity, and such segments outside the parallel parts overwhelmingly outnumber the parallel ones. Instead of averaging the similarity values, the similarity between the two documents can be also assessed by their structural alignment and the similarity of their parallel segments, visualized using graphs such as heatmaps and histograms (Figures 1, 2, and 3). Heatmaps, particularly those based on 20 lemmas, provide the most accurate depiction of similarity between the chunks in parallel form.

In the heatmap of MS.1.7 and KS.9.1 (Figure 2), the diagonal line highlighted in a light color indicates a high similarity of the chunks, illustrating that these two sections share parallels in the same order. Conversely, the heatmap of MS.1.6 and KS.8 (Figure 1) does not exhibit such close parallelism. The heatmap of MS.1.9 and KS.9.11 (Figure 3) shows a similar pattern to that of MS.1.7 and KS.9.1. The histograms corroborate these findings: MS.1.6 and KS.8 contain few sentences with a high similarity above 0.95, while MS.1.7 with KS.9.1 and MS.1.9 with KS.9.11 do. This suggests that MS.1.7 and MS.1.9 were composed under similar conditions, in close contact with KS, likely during a later period of composition.

4.2 Stylometry

The stylo package is used to perform cluster analysis and principal component analysis (PCA) on the texts divided into 20-lemma and 100-lemma chunks.

Figures 4 and 5 show the resulting dendrograms. The cluster analysis results align with our expectations for the known evaluation comparisons. MS.1.1 is consistently separated from the other texts, confirming its distinct nature. The pairs of parallel sections in MS and KS are correctly grouped together, indicating their stylistic similarity. Importantly, the PCA results (Figure 6) provide insights into our main focus, the comparison of MS.1.9 ↔ KS.9.11. This pair shows a closer stylistic relationship compared to MS.1.6 ↔ KS.8, but similar to MS.1.7 ↔ KS.9.1. This suggests that MS.1.9 and KS.9.11 likely share a similar compositional context or period with MS.1.7 and KS.9.1, supporting our hypothesis of their later period of composition and closer intertextual relationship. These stylometric results, particularly the PCA, complement our findings from word embeddings and text reuse detection, providing a multi-faceted view of the textual relationships in our corpus.

4.3 Text Reuse

The text reuse detection using TRACER yields the following number of parallels between the text pairs in Table 2.

Detection of MS.1.1 ↔ MS.1.6 reveals no reuse (parallel sentence), and those of other sections show a number of reuse, which indicates the analysis functions appropriately. The detection of 100-lemma corpora provides the number of close parallels. The highest number of parallels are found be-

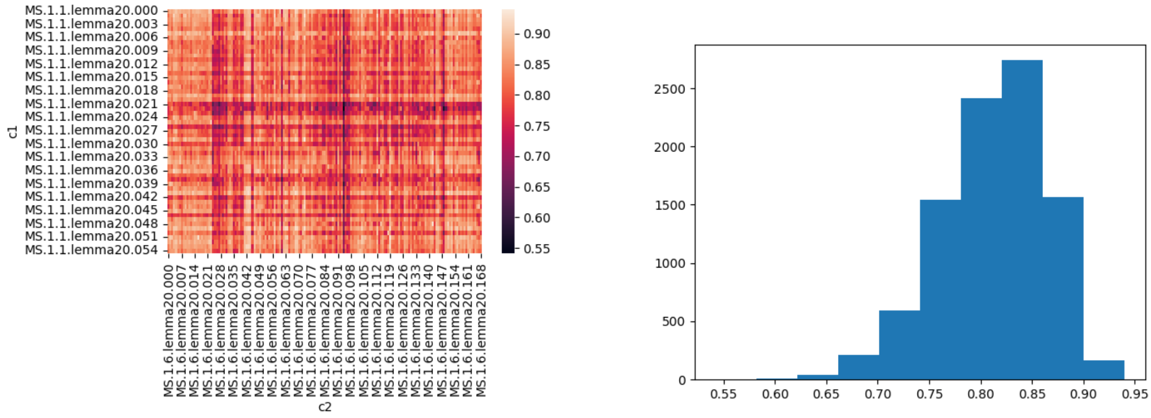

Figure 1: Word2Vec: heatmap and histogram of MS.1.6 ↔ KS.8 (20 lemma)

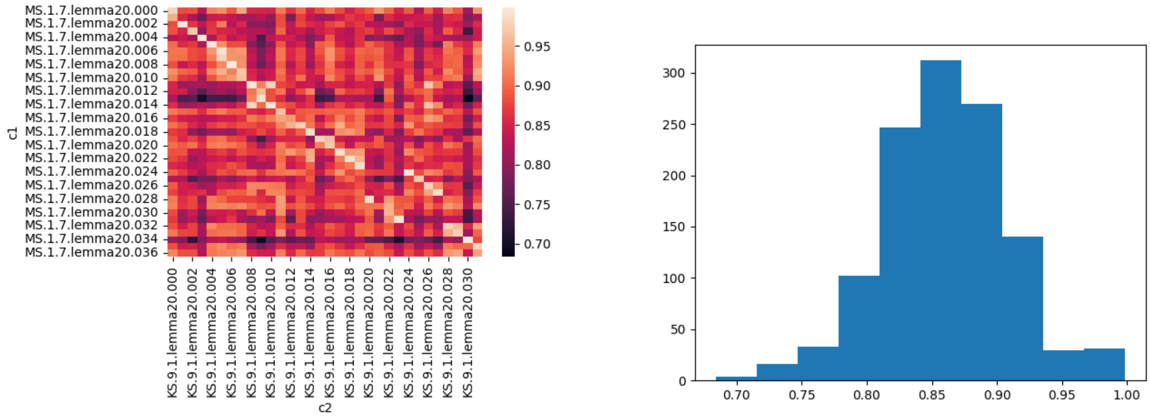

Figure 2: Word2Vec: heatmap and histogram of MS.1.7 ↔ KS.9.1 (20 lemma)

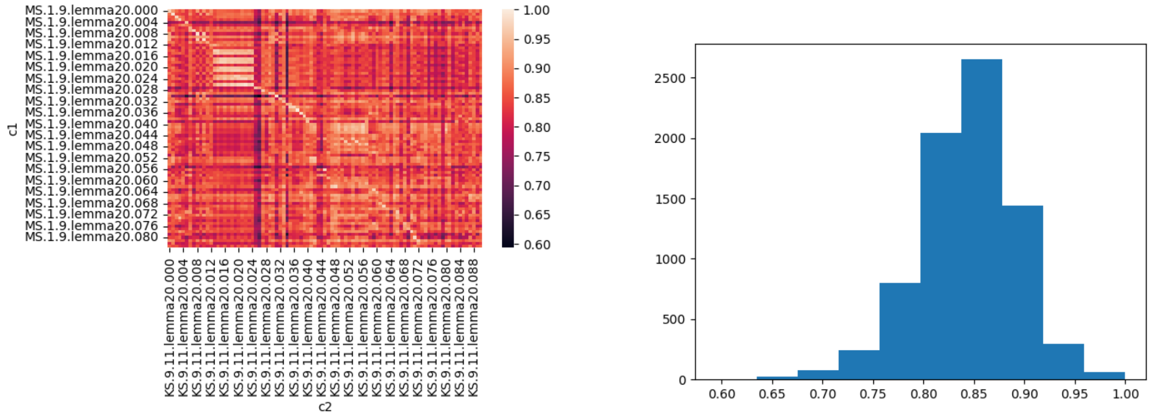

Figure 3: Word2Vec: heatmap and histogram of MS.1.9 ↔ KS.9.11 (20 lemma)

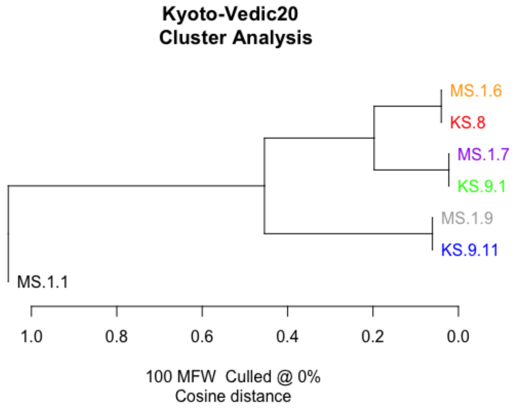

Figure 4: Cluster analysis of 20-lemma chunks using stylo

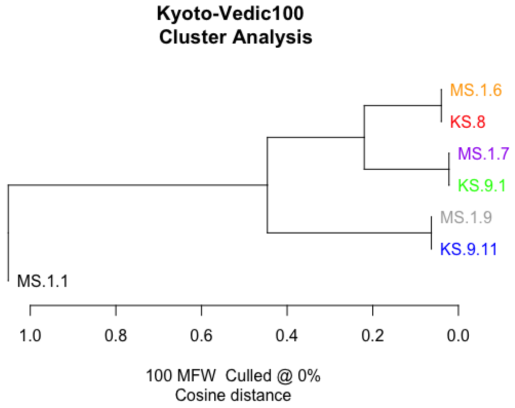

Figure 5: Cluster analysis of 100-lemma chunks using stylo

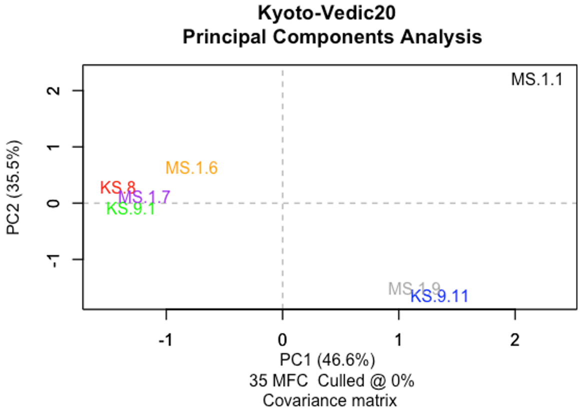

Figure 6: Stylo: Principal Components Analysis (20 lemma)

Table 2: Number of text reuse candidates detected by TRACER

Text Pair	20-lemma	100-lemma
MS.1.1 ↔ MS.1.6	N/A	N/A
MS.1.6 ↔ MS.1.7	13	3
MS.1.6 ↔ KS.8	8	15
MS.1.7 ↔ KS.9.1	55	10
MS.1.9 ↔ KS.9.11	209	15

tween MS.1.9 ↔ KS.9.11, followed by MS.1.7↔ KS.9.1. The detection of MS.1.6 ↔ KS.8 in 100-lemma corpus shows similar number of parallels to these two comparisons, which contradicts the previous study and of the analyses with Word2Vec and Stylo. The reason is that the size of the corpora is different (MS.1.6 contains 3816 words, MS.1.7 contains 819 words, MS.1.9 contains 1627 words). Due to the different sizes of the corpora, it is not appropriate to determine the similarity between sections based on the absolute number of parallels. However, graphs can compensate for this limitation.

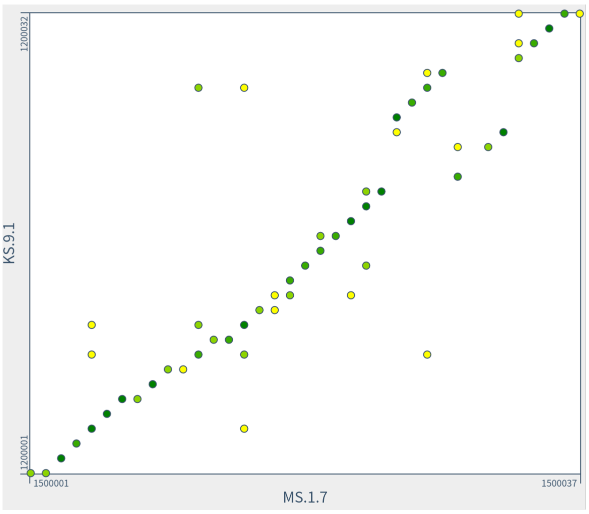

Figure 7: TRACER: MS.1.7 ↔ KS.9.1 (20 lemmas)

The graphs of MS.1.7 ↔ KS.9.1 (Figure 7) and MS.1.9 ↔ KS.9.11 (Figure 8) indicate the structural alignment, which is observed in the diagonal line of parallels, while that of MS.1.6 ↔ KS.8 (Figure 9) does not, as the heatmaps of Word2Vec indicated.

5 Conclusion

This paper presented an analysis of semantic similarity and text reuse in selected Vedic Sanskrit texts using word embedding, stylometric method and

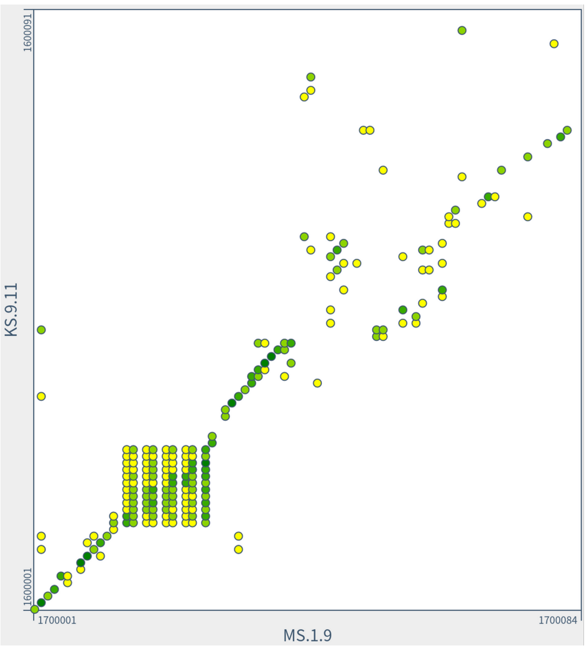

Figure 8: TRACER: MS.1.9 ↔ KS.9.11 (20 lemmas)

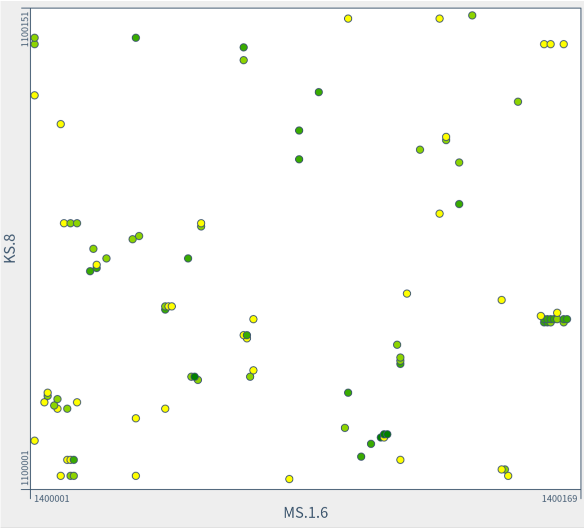

Figure 9: TRACER: MS.1.6 ↔ KS.8 (20 lemmas)

TRACER. The results from these approaches indicate patterns of similarity and clustering between different portions of the texts, which can be justified by previous literary studies. By analyzing the similarity of MS.1.9 ↔ KS.9.11, we inferred that MS.1.9 might share a similar historical period with MS.1.7.

- Using word embedding, the similarity between the pairs of sections was appropriately analyzed, the structural alignment was demonstrated well in the form of heatmap. The histograms helped us with understanding of similarity.
- The cluster analysis using stylo groups the corpora into intuitive clusters, with clearer separation at 100-lemma chunks.
- The text reuse detection using TRACER finds the highest number of parallels between MS.1.9↔KS.9.11 and MS.1.7↔KS.9.1, aligning with the stylometric clusters and the scores by the word embedding. The graphs show the structural alignment very well.

In conclusion, the computational analysis provides insights into the relationships between the texts and their sections in Vedic literature, which can clarify the process of its composition. The general agreement between the word embedding, text reuse detection and stylometric approach enhances the validity of the findings, and various visualizations of various analyses complemented each other's weaknesses and contributed to a more accurate understanding. Moreover, this study demonstrates that smaller chunk sizes are beneficial for finding parallels. On the other hand, documents with larger chunk sizes encompass various common topics. This means that documents with larger chunk sizes proportionally contain fewer topics that semantically distinguish them from each other, making it difficult to identify parallel relations with larger chunk sizes. Therefore, a smaller chunk size is more suitable for our purpose of finding parallels or verifying parallel relations between texts.

This research demonstrates the potential of computational methods in Vedic Sanskrit studies, and other ancient language corpora. Future work can extend the analysis to more texts, explore other embedding models and stylometric techniques, and closely examine the nature of the parallels identified. We hope this encourages further collaborative research at the intersection of indology and computational linguistics.

Limitations

While our study provides valuable insights into the similarity and intertextuality in Vedic Sanskrit literature using computational methods, it is important to acknowledge certain limitations. The analysis is based on a limited corpus size, focusing on selected sections from two Vedic texts, and the pre-processing of the texts relies on one possible interpretation, which could lead to variations in the results. The Word2Vec model used may not fully capture the semantic nuances and complexities of Vedic Sanskrit, and more advanced models like BERT were not explored due to the limited size of the training dataset. The chunk sizes used for analysis were chosen based on meaningful coherence, but different sizes may provide additional insights. The stylometric analysis focused primarily on cluster analysis and principal component analysis, while other techniques could reveal further stylistic patterns. The text reuse detection effectively identifies parallel passages, but their significance requires further qualitative analysis by domain experts. It is important to note that the computational methods used are complementary to traditional philological and linguistic analysis, and integration with existing studies is crucial for a holistic understanding. Despite these limitations, our research demonstrates the potential of computational approaches in studying ancient languages and texts, and further interdisciplinary collaborations and advancements in computational methods can greatly contribute to this field of study.

Ethics Statement

This research aims to advance the understanding of ancient Vedic Sanskrit texts through computational methods while adhering to ethical considerations. The computational analysis complements traditional approaches, and the interpretation of results requires the expertise of Indologists and Sanskrit scholars. We recognize the cultural and religious significance of the Vedic texts and approach the analysis with respect and sensitivity. The methods and tools used are open-source, promoting transparency and reproducibility. We acknowledge the risk of misinterpretation or oversimplification and emphasize the need for caution in drawing conclusions. This research has the potential to contribute to the preservation and understanding of ancient Indian heritage, inspiring further interdisciplinary research and public engagement. We are committed to conducting this research with integrity, transparency, and respect for the texts and the communities that hold them sacred.

References

Orna Almogi, Lena Dankin, Nachum Dershowitz, and Lior Wolf. 2019. A hackathon for classical Tibetan. *Journal of Data Mining & Digital Humanities*, (Towards a Digital Ecosystem: NLP. Corpus infrastructure. Methods for Retrieving Texts and Computing Text Similarities).

Kyoko Amano. 2009. *Maitrāyaṇī Saṃhitā I-II. Übersetzung der Prosapartien mit Kommentar zur Lexik und Syntax der älteren vedischen Prosa*, volume 9 of *Münchner Forschungen zur historischen Sprachwissenschaft*. Hempen Verlag, Bremen.

Kyoko Amano. 2014-2015. Zur Klärung der Sprachschichten in der Maitrāyaṇī Saṁhitā. *Journal of Indological Studies*, 26/27:1–36.

Kyoko Amano. 2020. What is 'knowledge' justifying a ritual action? uses of ya evaṁ veda / ya evaṁ vidvān in the Maitrāyaṇī Saṁhitā. In C. Redard, J. Ferrer-Losilla, H. Moein, and P. Swennen, editors, *Aux sources des liturgies indo-iraniennes*, volume 10 of *Collection Religions, Comparatisme - Histoire - Anthropologie*, pages 39–68. Presses Universitaires de Liège, Liège.

Marco Büchler. 2013. Informationstechnische Aspekte des Historical Text Re-use.

Marco Büchler, Philip R. Burns, Martin Müller, Emily Franzini, and Greta Franzini. 2014. Towards a historical text re-use detection. In Michael W. Berry and Jacob Kogan, editors, *Text Mining*, pages 221–238. Springer.

Marco Büchler, Greta Franzini, Emily Franzini, Maria Moritz, and Kirill Bulert. 2018. TRACER-a multilevel framework for historical text reuse detection.

Marco Büchler, Annette Geßner, Gerhard Heyer, and Thomas Eckart. 2010. Detection of citations and textual reuse on ancient greek texts and its applications in the classical studies: eAQUA project. In *Proceedings of Digital Humanities 2010*, pages 113–114.

Marco Buechler, Gerhard Heyer, and Sabine Gründer. 2008. eAQUA– bringing modern text mining approaches to two thousand years old ancient texts. In *Proceedings of e-Humanities–An Emerging Discipline, workshop at the 4th IEEE International Conference on e-Science*.

Neil Coffee, Jean-Pierre Koenig, Shakthi Poornima, Christopher W Forstall, Roelant Ossewaarde, and

Sarah L Jacobson. 2012. The tesserae project: intertextual analysis of latin poetry. *Literary and linguistic computing*, 28(2):221–228.

Maciej Eder, Jan Rybicki, and Mike Kestemont. 2016. Stylometry with R: A Package for Computational Text Analysis. *The R Journal*, 8(1):107–121.

Greta Franzini, Mike Kestemont, Gabriela Rotari, Melina Jander, Jeremi K Ochab, Emily Franzini, Joanna Byszuk, and Jan Rybicki. 2018a. Attributing authorship in the noisy digitized correspondence of jacob and wilhelm grimm. *Frontiers in Digital Humanities*, 5:4.

Greta Franzini, Marco Passarotti, Maria Moritz, and Marco Büchler. 2018b. Using and evaluating TRACER for an index fontium computatus of the summa contra gentiles of thomas aquinas. In Alessandro Mazzei Elena Cabrio and Fabio Tamburini, editors, *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it) 2018: Torino, Italy, December 10–12*, pages 199–205.

Oliver Hellwig and Sebastian Nehrdich. 2018. Sanskrit word segmentation using character-level recurrent and convolutional neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2754–2763, Brussels, Belgium. Association for Computational Linguistics.

Oliver Hellwig, Sebastian Nehrdich, and Sven Sellmer. 2023. Data-driven dependency parsing of vedic sanskrit. *Language Resources and Evaluation*, 57(3):1173–1206.

Oliver Hellwig, Salvatore Scarlata, Elia Ackermann, and Paul Widmer. 2020. The treebank of vedic Sanskrit. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5137–5146, Marseille, France. European Language Resources Association.

Amrith Krishna, Vishnu Sharma, Bishal Santra, Aishik Chakraborty, Pavankumar Satuluri, and Pawan Goyal. 2019. Poetry to prose conversion in Sanskrit as a linearisation task: A case for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1160–1166, Florence, Italy. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

So Miyagawa. 2021. Digitization of Coptic manuscripts and digital humanities: Tools and methods for Coptic studies. *The International Journal of Levant Studies*, 2:29–61.

So Miyagawa. 2022. *Shenoute, Besa and the Bible Digital Text Reuse Analysis of Selected Monastic Writings from Egypt*. SUB Göttingen.

So Miyagawa, Amir Zeldes, Marco Büchler, Heike Behlmer, and Troy Griffitts. 2018. Building linguistically and intertextually tagged coptic corpora with open source tools. In Chikahiko Suzuki, editor, *Proceedings of the 8th Conference of Japanese Association for Digital Humanities*, pages 139–41. Center for Open Data in the Humanities.

Matteo Romanello and Simon Hengchen. 2021. Detecting text reuse with passim. *Programming Historian*, 10.

Justin Anthony Stover, Yaron Winter, Moshe Koppel, and Mike Kestemont. 2016. Computational authorship verification method attributes a new work to a major 2nd century African author. *Journal of the Association for Information Science and Technology*, 67(1):239–242.