Journal of Data Mining and Digital Humanities

EPI sciences overlay journals

Computational Pathways to Intertextuality of the Ancient Indian Literature: A Multi-Method Analysis of the Maitrāyaṇī and Kāṭhaka Saṃhitās

So Miyagawa (University of Tsukuba), Yuzuki Tsukagoshi (University of Tokyo), Yuki Kyogoku (Leipzig University), and Kyoko Amano (Kyoto University)

Abstract

This paper investigates semantic similarity and intertextuality in selected texts from the Vedic Sanskrit corpus, in particular in the Maitrāyaṇī Saṃhitā (MS; Amano, 2009) and Kāṭhaka Saṃhitā (KS). Three calculation methods are used: Word2Vec for word embeddings, the package `stylo` for stylometric analysis and TRACER for text reuse detection. By comparing different sections of text with different granularity, similarity patterns and structural matches are uncovered that provide information about text relationships and chronology. Word embeddings capture semantic similarities, while stylometric analysis uncovers clusters that distinguish the texts from one another. TRACER identifies parallel passages that indicate likely instances of text reuse. Our multi-method analysis confirms previous philological studies and suggests that MS.1.9 is consistent with later redactional layers, similar to MS.1.7. The results emphasize the potential of computational methods in the study of ancient Sanskrit literature to complement traditional approaches and emphasize that intertextual parallels can be better identified with smaller data sets. These approaches extend the methodological boundaries of Indology and point to new research avenues for analyzing ancient texts.[1]

Keywords

Vedic; stylometry; word embedding; text reuse detection; sandhi; similarity measurement

I INTRODUCTION

The study of Vedic Sanskrit literature preserves invaluable cultural and historical information from ancient India. However, these texts show distinct challenges due to their linguistic complexity and modes of composition/transmission. Recently, computational methods have offered promising new opportunities for studying such texts on a large scale.

In this work, we focus on the Maitrāyaṇī Saṃhitā (MS) and the Kāṭhaka Saṃhitā (KS), which are traditionally dated to about 900–700 BCE and comprise the oldest prose portions of Sanskrit texts. While previous philological studies have already identified parallels between specific sections of these texts, as early as Schroeder [1881–1886], the specific degree and patterns of

[1]This paper is a revised and extended version of the following paper:

So Miyagawa, Yuki Kyogoku, Yuzuki Tsukagoshi, and Kyoko Amano. 2024. Exploring Similarity Measures and Intertextuality in Vedic Sanskrit Literature. In Mika Hämäläinen, Emily Öhman, So Miyagawa, Khalid Alnajjar, and Yuri Bizzoni (eds.), *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 123–131, Miami, USA. Association for Computational Linguistics.DOI: 10.18653/v1/2024.nlp4dh-1.12

similarity in different chapters have been less systematically studied. In particular, older layers often show fewer word-level correspondences; later sections show strong lexical convergences, and editorial syncretism.

Although comprehensive research on the entire MS and KS must wait due to insufficient data, within the range for which data are available, we investigate whether MS.1.9–KS.9.11 displays a degree of intertextual synergy that allows us to determine its genesis, similar to "late-phase" pairs such as MS.1.7–KS.9.1, or whether it is closer to "early-phase" pairs such as MS.1.6–KS.8. While previous studies have not analyzed the similarity of MS.1.9–KS.9.11, our comparative analysis may newly reveal their degree of similarity and allow us to estimate their relative chronology. We use three different computational approaches:

- Word embeddings (Word2Vec) for semantic similarity
- Stylometric analysis using the `stylo` package
- Text reuse detection using TRACER

By comparing these approaches, we get a more comprehensive view of textual relationships and can determine whether MS.1.9 is more consistent with later editorial layers or if it corresponds to an earlier layer. In addition, we investigate how block sizes (20 vs. 100 or 200 lemmas) affect similarity detection and text reuse.

II HISTORICAL AND PHILOLOGICAL FOUNDATIONS

2.1 Early Indological Efforts and the Emergence of Layered Textual Models

In the 19th and 20th centuries, studies of the Vedas examined the relationships between different Vedic texts and proposed relative chronologies. However, while in the case of the Rigveda differences in the editorial layers within individual texts were recognized, this perspective was rarely applied to other Vedic texts.

Amano's 2014—2015 research was the first to establish that multiple linguistic layers with different characteristics exist within the MS [Amano, 2014–2015]. Furthermore, it was shown that the relationship between the MS and its parallel text, the KS, has evolved, with these changes reflected in the varying degrees of similarity between the two texts. Although quantitative methods were used in this study, no computational tools were used, so digital approaches allow for more precise similarity calculations and more comprehensive analyses.

2.2 Transition to Digital Corpora and Lexical Databases

The digitization of Sanskrit materials and the creation of lexical databases have revolutionized Indological research. Tools such as the Digital Corpus of Sanskrit (`DCS`) and neural morphological analyzers [Hellwig et al., 2020] have enabled large-scale comparisons that were unthinkable in a strictly manual environment. Particularly for Vedic Sanskrit, morphological complexity and sandhi phenomena pose significant challenges, but new NLP models and machine-readable corpora support robust text comparisons.

Recent computational philology efforts have provided stylometric frameworks and text reuse detection that help confirm or challenge earlier philological hypotheses. Our study builds upon such infrastructure, using cleaned, un-sandhied, and lemmatized corpora to ensure consistent textual segmentation.

2.3 Related Computational Work

For Sanskrit NLP development, Hellwig and Nehrdich [2018] created a Vedic treebank, while Hellwig et al. [2023] introduced a dependency parser for Ṛgvedic Sanskrit. Krishna et al. [2019] demonstrated deep-learning-based analysis for Sanskrit poetic style. For stylometry in classical literature, Stover et al. [2016] successfully attributed a classical text to Apuleius using the `stylo` package [Eder et al., 2016].

Text reuse detection in ancient languages has seen significant development via TRACER [Büchler, 2013, Büchler et al., 2018], which has been applied to Greek [Büchler et al., 2010], Latin [Franzini et al., 2018], Coptic [Miyagawa, 2021, Miyagawa et al., 2018, Miyagawa, 2022], Tibetan [Almogi et al., 2019], and others. Given Sanskrit's morphological complexity, TRACER is adaptable through custom lemma and synonym files, making it a promising tool for exploring parallels in Vedic literature.

III CORPUS SELECTION RATIONALE AND PREPROCESSING

The texts selected for this study come from carefully chosen MS and KS segments. Each of the seven segments examined contains unique thematic and ritual content, allowing us to explore varying degrees of editorial overlap and synergy. Segments such as MS.1.1 (1145 words[2]) address new and full-moon sacrifice formulas, while MS.1.6 (3816 words) and MS.1.7 (819 words) focuses on establishing and reestablishing sacred fires, respectively. These latter two segments are particularly relevant for comparison with parallel sections in KS, namely KS.8 (3519 words) and KS.9.1 (818 words). Additionally, MS.1.9 (1627 words) and KS.9.11 (1721 words) comprise overlapping materials concerning secret spells, although the extent of their parallels remains relatively underexplored.

- **MS.1.1** (1145 words): Contains new and full moon sacrifice formulas
- **MS.1.6** (3816 words): Explanation of establishing sacred fires (old editorial phase; parallel is KS.8)
- **MS.1.7** (819 words): Explanation of reestablishing sacred fires (later editorial phase; parallel is KS.9.1)
- **MS.1.9** (1627 words): Explanation of secret spells (parallel is KS.9.11)
- **KS.8** (3519 words): parallels MS.1.6
- **KS.9.1** (818 words): parallels MS.1.7
- **KS.9.11** (1721 words): parallels MS.1.9

By selecting these segments, our corpus highlights the older and younger strata within MS and KS. For instance, MS.1.1 is markedly different from MS.1.6 or MS.1.7 regarding philological features and ritual focus, so we anticipate minimal similarity in editorial style and content. In contrast, MS.1.6 and MS.1.7 share content regarding sacred fires and thus may exhibit higher similarity scores than MS.1.1 vs. MS.1.6. Parallel segments across MS and KS are expected to show somewhat high similarity. KS.8 parallels MS.1.6, and KS.9.1 parallels MS.1.7. These two pairs present different aspects. That is, the former was established in an earlier period, while the latter was formed through borrowing between the two in the later stages of editing. The relatively unexamined pairing MS.1.9 vs. KS.9.11 invites an open question: does their relationship resemble older parallels (less similarity) or align more with late-phase synergy (more significant similarity)?

To organize our inquiry, we focus on five main cross-comparisons: (1) MS.1.1 vs. MS.1.6, (2)

[2] A compound consisting of two words was not considered as one word but as two words.

MS.1.6 vs. MS.1.7, (3) MS.1.6 vs. KS.8, (4) MS.1.7 vs. KS.9.1, and (5) MS.1.9 vs. KS.9.11. These targeted comparisons offer a structured approach for examining the degree of textual reuse, editorial consistency, and semantic coherence across passages that vary in ritual function and presumed chronological layering. Furthermore, this multi-textual framework is designed to elucidate whether certain textual resemblances stem from genuine borrowing, or shared underlying traditions.

Philologically, MS.1.1 differs significantly from MS.1.6 or MS.1.7, so a low similarity is expected. Conversely, MS.1.6 vs. MS.1.7 share content, so a higher similarity is expected. Regarding MS-KS comparisons, older pairs like MS.1.6–KS.8 appear less similar, whereas MS.1.7–KS.9.1 show strong synergy. MS.1.9–KS.9.11 has been relatively unexamined; the question is whether it patterns with older or later pairs.

All texts undergo a multi-step preprocessing:

1. **Un-sandhi**: We remove sandhi to ensure reliable word segmentation (cf. Hellwig et al. 2020).
2. **Lemmatization**: Each word is mapped to its canonical root form.
3. **Chunking**: We create segments at the section level and fixed-size segments (20, 100, 200 lemmas).

All text data goes through a three-step preprocessing pipeline to ensure consistent comparability across all corpora. First, we perform un-sandhi operations, removing external sandhi and enabling a more standardized word-level segmentation [Hellwig et al., 2020]. This step is crucial in a highly inflectional language like Sanskrit, as sandhi can otherwise blur the boundaries between words. Second, we apply lemmatization, mapping each token to its canonical root form. Lemmatization mitigates the effects of morphological variation and allows us to detect parallel vocabulary even when case endings, verb conjugations, or other morphological features differ. Third, we apply a chunking strategy that divides each text into thematically oriented chunks and segments of fixed size (20, 100, and 200 lemmas). Smaller chunks reveal micro-level parallels—such as short repeated formulas—while larger chunks detect similarities that are precisely aligned across the entire chunk.

In summary, our corpus selection reflects a conscious effort to capture older and newer editorial layers within Vedic ritual traditions, while our preprocessing steps maximize accuracy and consistency in textual comparisons. By studying a range of cross-comparisons and segment sizes, we seek to shed light on the complexity of the Vedic textual transmission and distinguish between direct borrowing, convergent textual development, and more general shared features of the ritual tradition.

IV METHODOLOGICAL FRAMEWORK: THREE PILLARS

4.1 Semantic Embeddings (Word2Vec)

Word2Vec [Mikolov et al., 2013] provides a powerful method for capturing semantic relationships by learning continuous vector representations of words. In this study, we trained the model on a broad Vedic Sanskrit corpus, carefully excluding MS and KS so that the learned representations are not biased toward the directly studied texts. This approach aims to establish a neutral semantic "baseline": all similarities detected within MS and KS should reflect true parallels, rather than represent a circular reinforcement of pre-learned vectors from the same data.

After training, each text block is represented as a "document vector" created by averaging the embeddings of all words within that block. Cosine similarity measures the degree of overlap between these vectors and allows one to detect semantic parallels in otherwise distinct text segments. Smaller block sizes (e.g., 20 lemmas) are useful when looking for short formulaic parallels, as they help isolate repeated motifs or phrases at a fine-grained level. Larger block sizes detect similarities when there are similarities across the entire block, i.e., there are strong structural similarities. In other words, they can obscure micro-level alignments. Balancing these window sizes allows for a multi-resolution perspective: one can zoom in on short parallel expressions while detecting longer structural alignments that span multiple lines.

4.2 Stylometry (`stylo` Package)

Stylometry focuses on the statistical analysis of common words and function words to uncover underlying writing styles or editorial signatures. It provides a complementary dimension to semantic embedding by shifting attention away from topical or lexical meaning and towards habitual linguistic patterns. In this project, we relied on the `stylo` R package [Eder et al., 2016] to perform cluster analysis and principal component analysis (PCA). By quantifying the relative frequency of frequently occurring elements (e.g. pronouns, particles or prepositions), stylometry can reveal regularities that authors or editors may not consciously control.

With the help of PCA, text segments with similar stylistic features are grouped together in a multidimensional space defined by these frequently used words. If two passages of MS or KS exhibit intense proximity in PCA space, it suggests a shared editorial approach or standardized phraseology.

The combination of stylometry and Word2Vec-based semantics can allow us to distinguish whether segments are linked primarily on the basis of content (semantic embedding) or form (stylistic markers). This dual perspective is particularly valuable in ancient textual traditions, where repeated formulas or standardized language use can reflect conscious editorial decisions rather than simple duplication of meaning.

4.3 Text Reuse Detection (TRACER)

Text reuse detection methods, as implemented in TRACER [Büchler, 2013, Büchler et al., 2018], target literal or near literal overlaps, complementing the broader focus of semantic embeddings and stylometry. Instead of analyzing general stylistic or semantic fields, TRACER searches for exact or near-exact word sequences. This approach is particularly well suited to detecting parallel sentences or entire formulaic lines—a hallmark of Vedic textual traditions, which often rely on precise repetition for ritual or mnemonic purposes.

Visualizing these matches (see figures 6 and 7) in a grid helps to confirm whether parallel passages line up in a diagonal pattern, indicating similar sequences or direct textual borrowings. If the overlaps are dense, this strongly indicates editorial synergy through deliberate borrowing between MS and KS. In contrast to purely semantic methods, TRACER's sensitivity to the identity of literal strings underscores the importance of recognizing literal borrowings. The integration of TRACER with Word2Vec and stylometry thus provides a multi-pronged methodology: semantic embeddings capture thematic resonance, stylometry uncovers stylistic congruence, and text reuse detection flags direct repetitions and formulaic echoes. This tri-fold approach provides a robust framework for investigating how editorial processes shape complex ancient corpora.

V RESULTS

The results of our three-way analysis are shown at Table 1 for semantic embeddings, Figures 1 and 2 for stylometry, and Tables 2 and Figures 6 and 7 for text reuse analysis.

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

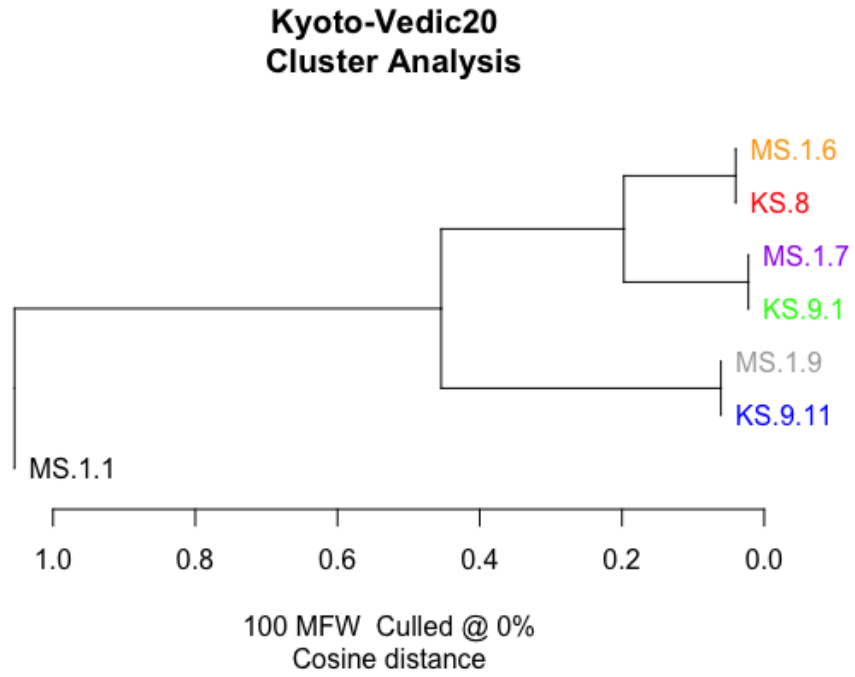

Figure 1: Cluster analysis of 20-lemma chunks using stylo

5.1 MS.1.1 vs. MS.1.6 and MS.1.6 vs. MS.1.7

We first check whether the methods reflect the known difference between MS.1.1 and MS.1.6 as well as the known similarity between MS.1.6 and MS.1.7. The average Word2Vec similarity is indeed lower for MS.1.1-MS.1.6 and higher for MS.1.6–MS.1.7. Stylometry places MS.1.1 in a clear cluster, while MS.1.6 and MS.1.7 show partial convergence. TRACER provides no strong parallels for MS.1.1–MS.1.6, but finds several direct text matches for MS.1.6–MS.1.7. This consistency confirms that the tools behave as expected.

5.2 MS.1.6 vs. KS.8 and MS.1.7 vs. KS.9.1

Previously, Amano [2014–2015] and Amano [2020] found that MS.1.6–KS.8 (Fig. 3) share content but show relatively little direct reuse. In contrast, MS.1.7–KS.9.1 (Fig. 4) reflects a more integrated editing process that has many nearly identical lines. Our computational results agree: Word2Vec heatmaps show a robust diagonal pattern for MS.1.7–KS.9.1, stylometry

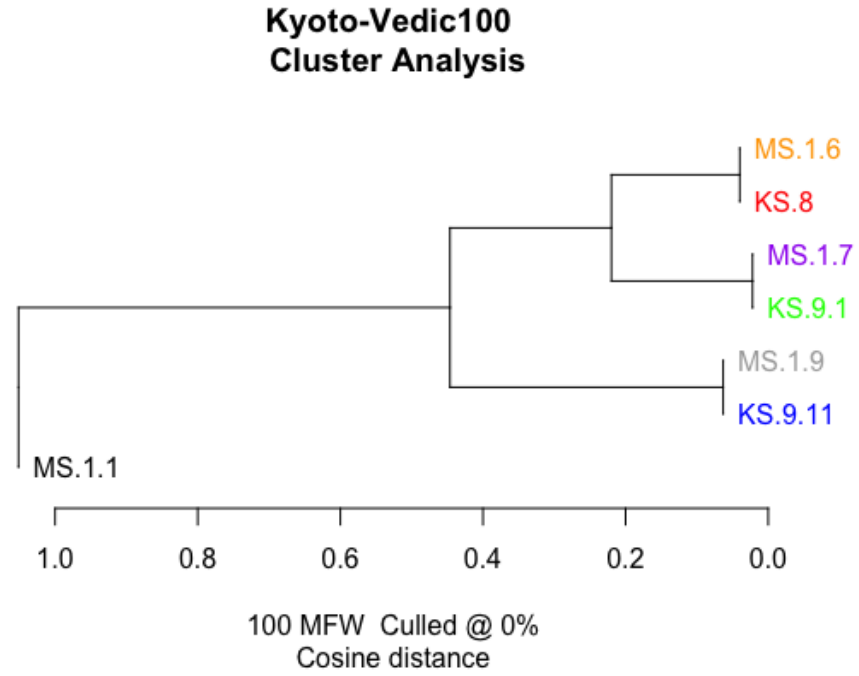

Figure 2: Cluster analysis of 100-lemma chunks using stylo

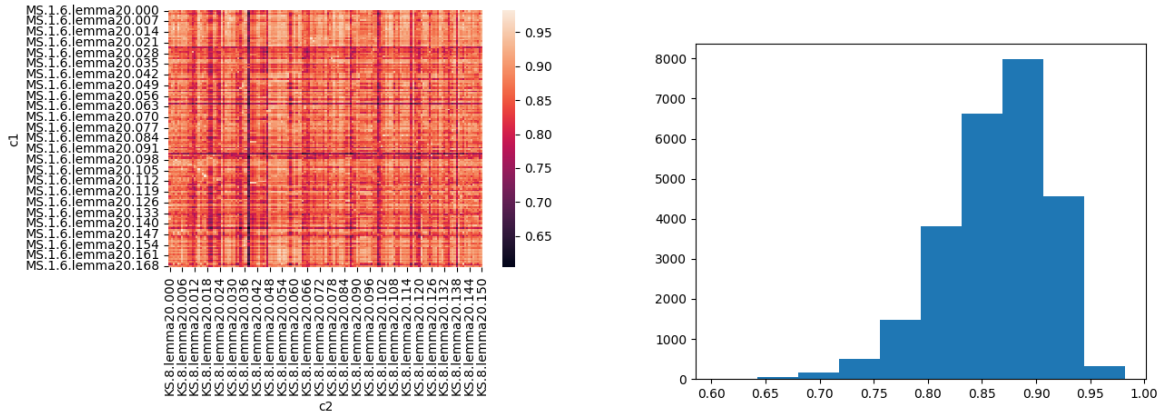

Figure 3: Word2Vec: heatmap and histogram of MS.1.6 ↔ KS.8 (20 lemma)

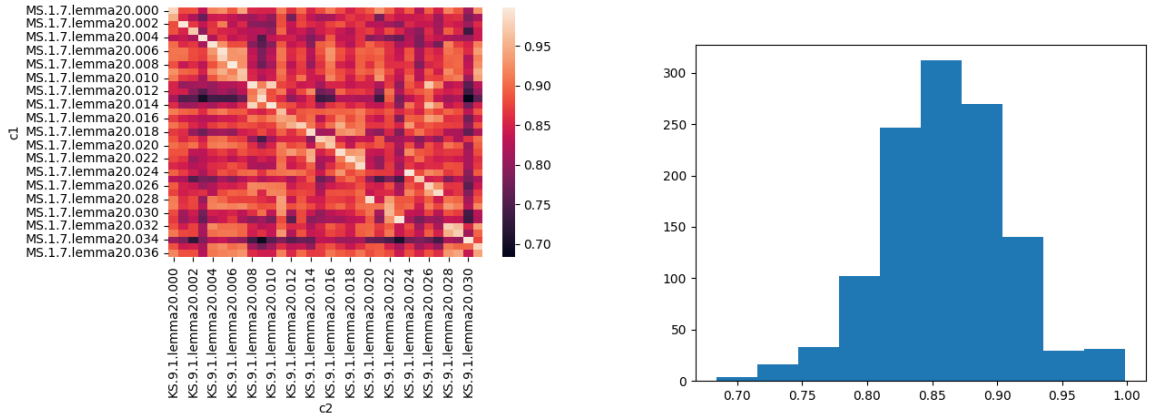

Figure 4: Word2Vec: heatmap and histogram of MS.1.7 ↔ KS.9.1 (20 lemma)

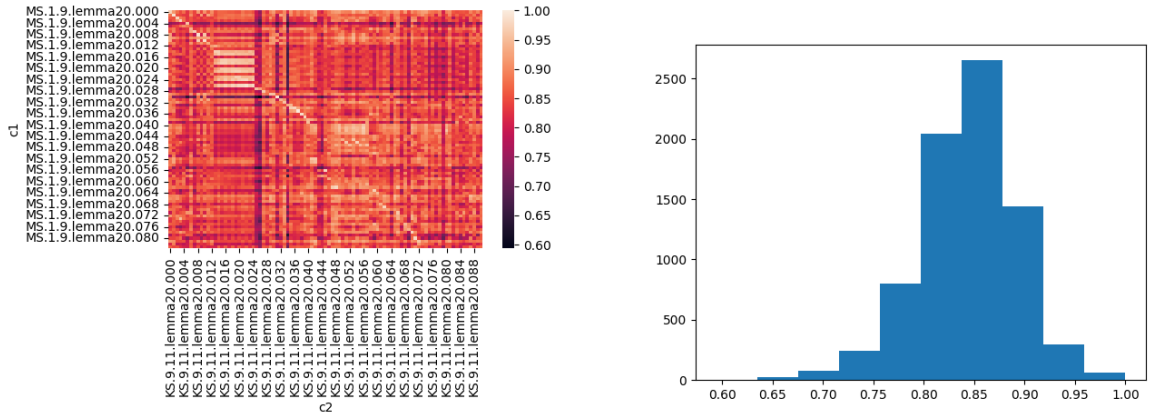

Figure 5: Word2Vec: heatmap and histogram of MS.1.9 ↔ KS.9.11 (20 lemma)

groups them together, and TRACER identifies numerous matching segments. MS.1.6–KS.8, on the other hand, shows fewer direct parallels and lacks a strong diagonal in the visualizations of text reuse.

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

VI MS.1.9 VS. KS.9.11: MAIN PUZZLE

The key question is whether MS.1.9 patterns have older or later editorial phases. Our data (Fig. 5) show:

1. **Word2Vec Heatmap & Histogram:** Diagonal alignment and substantial high-similarity chunks, similar to MS.1.7–KS.9.1.
2. **Stylometry (Cluster/PCA):** MS.1.9 clusters more closely with KS.9.11 than MS.1.6 does with KS.8.
3. **TRACER:** Large numbers of parallel strings, including some long matching phrases strongly suggest direct textual reuse.

Hence, MS.1.9 mirrors the synergy that characterizes late-phase text pairs. This implies that the relevant editorial processes likely occurred during a period where authors intended to standardize or unify textual tradition across multiple Saṃhitās.

VII DISCUSSIONS

Our findings shed light on an evolving editorial context in which authors increasingly borrowed and standardized textual material. This phenomenon is thought to have occurred as networks between communities became more closely connected over time, probably due to the expansion of routes and pathways in those days.

Despite remarkable methodological advances, our study is subject to inherent limitations. First, the corpus we analyzed is relatively small, with only seven text segments from the MS and the

KS. A larger data set would increase our statistical confidence. In order to expand the analysis data, we need to further advance the lemmatization of the original text. However, automated resolution of sandhi and lemmatization are not perfect, requiring extensive time for human corrections. Nevertheless, it is expected that this problem will be significantly improved by the new tool recently introduced by [Nehrdich et al., 2024]. Second, similarity scores may be sensitive to the choice of parameters—such as the size of Word2Vec windows or TRACER thresholds for match length—although convergent results from multiple methods increase confidence in the central results.

Future research would benefit from expanding the corpus and analyzing segments of the Vedic texts that have not yet been sufficiently studied, and would provide more detailed insights into editorial processes. Linking computational philology with philological, archaeological, or anthropological findings could show how textual changes correlate with historical changes in ritual practice.

VIII CONCLUSION

By applying Word2Vec embeddings, stylometric analysis, and TRACER-based text reuse detection to selected segments of MS and KS, we confirm that MS.1.9 is closely aligned with later editorial sections—mirroring the synergy seen in MS.1.7–KS.9.1. This strongly suggests that MS.1.9 was shaped under similar conditions of editorial homogenization, thereby supporting the hypothesis of its later-phase composition.

Moreover, we highlight that each method—semantic embeddings, stylometry, and text reuse detection—offers distinct but complementary insights. The consistent clustering of MS.1.9–KS.9.11 across approaches provides robust corroboration. Equally important, the "heatmap," "histogram," and "TRACER alignment" visualizations reveal that smaller chunk sizes (e.g., 20 lemmas) can be more sensitive to textual parallels than larger ones.

These results demonstrate the promise of computational methods in the analysis of ancient texts. Far from replacing philological scholarship, they expand our evidential base and refine the precision of interpretation. As more and more Vedic texts are digitized and advanced NLP tools develop, such integrative approaches will reshape our understanding of how these venerable texts were created, disseminated, and consolidated into the canons we inherit today.

Advancing the digital humanities framework for Vedic studies also requires a robust infrastructure. Standardized annotation formats, version-controlled repositories, and curated lexical databases would ensure reproducibility and encourage scholarly collaboration. Advances in optical character recognition (OCR) for Sanskrit manuscripts will further expand the availability of high-quality digital data. In parallel, community building efforts—such as workshops training Indologists in NLP and computational linguists in philology—will promote sustainable growth in this interdisciplinary field.

References

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

Kyoko Amano. *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, 2009.

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

Kyoko Amano. What is 'knowledge' justifying a ritual action? Uses of ya evaṃ veda / ya evaṃ vidvān in the Maitrāyaṇī Saṃhitā. 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, 2020.

Marco Büchler. *Informationstechnische Aspekte des Historical Text Re-use*. Phd thesis, 2013.

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

Marco Büchler, Annette Geßner, Gerhard Heyer, and Thomas Eckart. 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, 2010.

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

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

Greta Franzini, Marco Passarotti, Maria Moritz, and Marco Büchler. 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. 2018. URL http://ceur-ws.org/Vol-2253/paper22.pdf.

Oliver Hellwig and Sebastian Nehrdich. Sanskrit word segmentation using character-level recurrent and convolutional neural networks. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2754–2763, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1295. URL https://aclanthology.org/D18-1295.

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

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

Amrith Krishna, Vishnu Sharma, Bishal Santra, Aishik Chakraborty, Pavankumar Satuluri, and Pawan Goyal. Poetry to prose conversion in Sanskrit as a linearisation task: A case for low-resource languages. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1160–1166, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1111. URL https://aclanthology.org/P19-1111.

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

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

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

So Miyagawa, Amir Zeldes, Marco Büchler, Heike Behlmer, and Troy Griffitts. 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 139–141. Center for Open Data in the Humanities, Tokyo, 2018.

Sebastian Nehrdich, Oliver Hellwig, and Kurt Keutzer. One model is all you need: ByT5-Sanskrit, a unified model for Sanskrit NLP tasks. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13742–13751, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.805. URL https://aclanthology.org/2024.findings-emnlp.805/.

Leopold von Schroeder. *Maitrāyaṇī Saṃhitā*. Brockhaus, 1881–1886. Published in 4 volumes: 1881, 1883, 1885, 1886. Reprinted by Steiner, Wiesbaden, 1970.

Justin Anthony Stover, Yaron Winter, Moshe Koppel, and Mike Kestemont. 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, 2016.

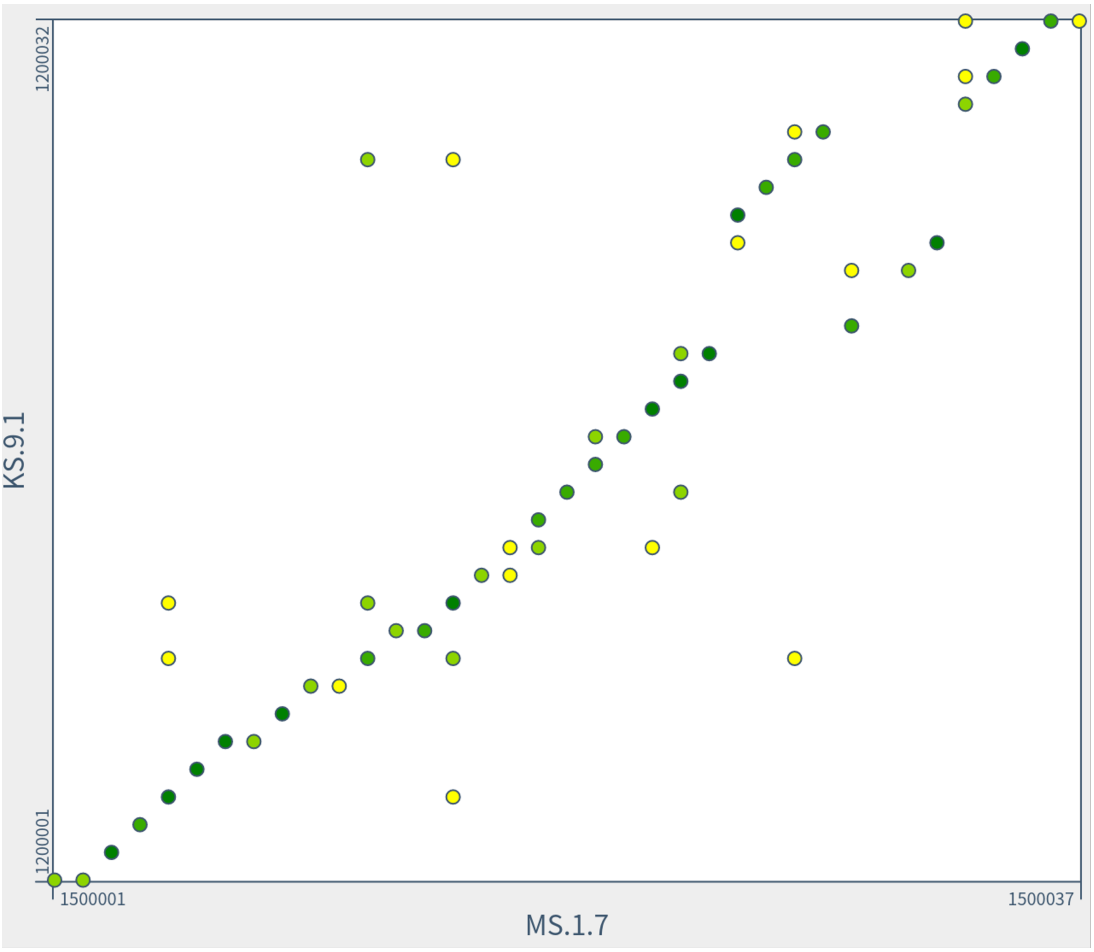

Figure 6: Text reuse detection between MS.1.7 and KS.9.1

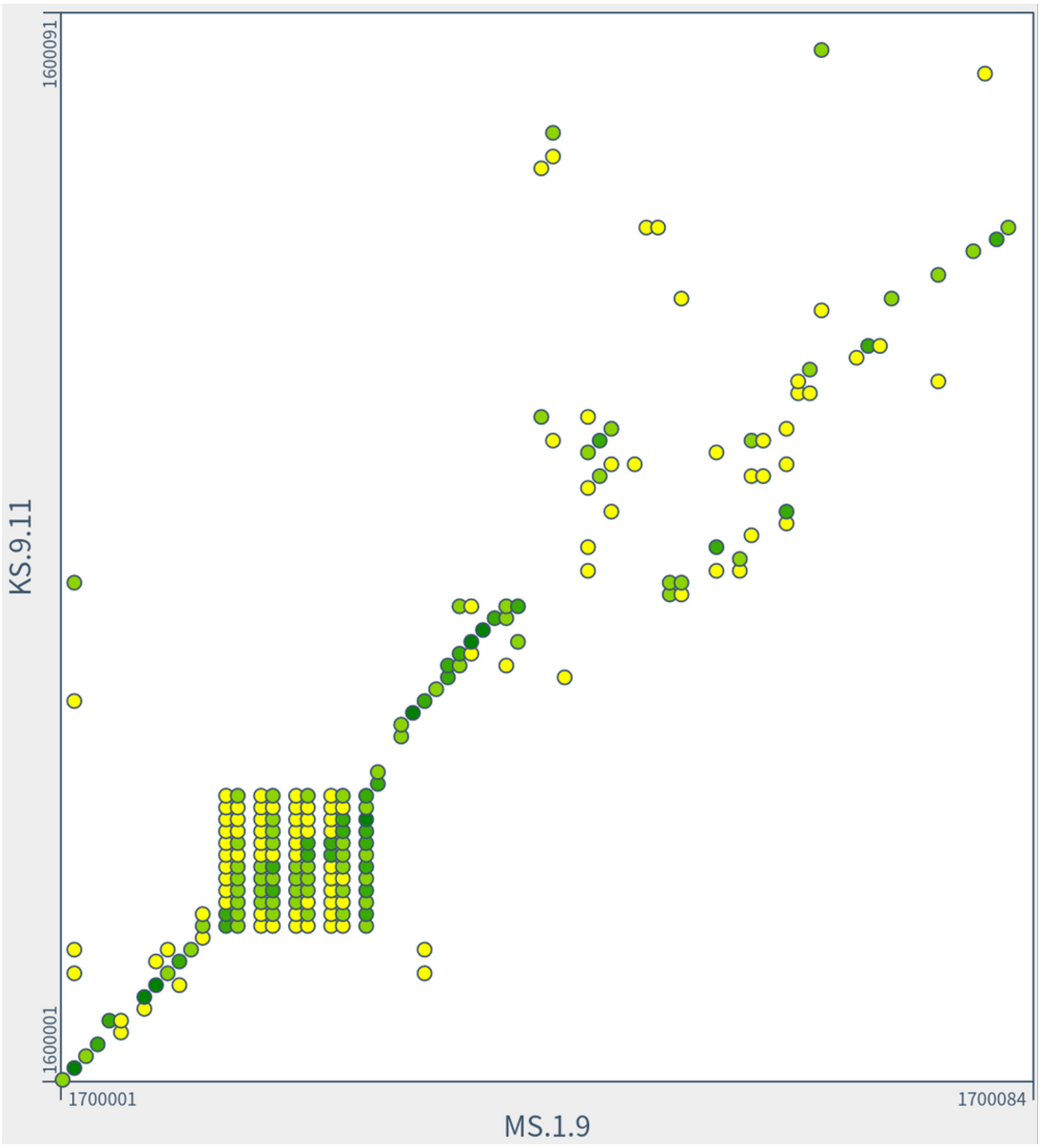

Figure 7: Text reuse detection between MS.1.9 and KS.9.11