Quantifying Text Reuse Across Three Kṛṣṇa Yajurveda Recensions Using Multi-Algorithm Computational Collation

So Miyagawa[1], Kyoko Amano[2], Yuzuki Tsukagoshi[3], Yuki Kyogoku[4]

[1]Institute of Humanities and Social Sciences, University of Tsukuba, Japan,
[2]Graduate School of Letters, Kyoto University, Japan,
[3]Graduate School of Humanities and Sociology, The University of Tokyo, Japan,
[4]Institute for South and Central Asian Studies, Leipzig University, Germany

Correspondence: miyagawa.so.kb@u.tsukuba.ac.jp

Abstract

The Kṛṣṇa Yajurveda survives in multiple recensions that share substantial ritual content, yet the degree and distribution of textual overlap across recensions have never been quantified systematically. This paper presents a computational analysis of text reuse across three recensions—the Maitrāyaṇī Saṃhitā (MS), the Kāṭhaka Saṃhitā (KS), and the Taittirīya Saṃhitā (TS)—for two ritual sections (Agnyupasthāna and Punarādhāna), using ICoMa (Intertextuality Collation Machine), a new web-based multi-algorithm collation tool. Five independent similarity algorithms consistently rank MS–KS as the most closely related pair, corroborating the philological consensus. Crucially, the two ritual sections exhibit strikingly different reuse profiles: Punarādhāna shows near-identical MS–KS overlap (up to 93.5%) with sharp divergence from TS, while Agnyupasthāna displays moderate, broadly distributed similarity across all three pairs. These contrasting patterns provide quantitative evidence that different ritual categories followed distinct paths of textual transmission within the Yajurvedic tradition. Thus, a way has been opened to explain the differences in similarity as reflecting differences in the periods at which the texts became fixed. ICoMa and the experimental data are freely available.

1 Introduction: The Kṛṣṇa Yajurveda and Its Recensions

The Vedas of ancient India constitute a vast corpus of literature centered on the religious rituals of their time. The oldest of these texts is the Ṛgveda, generally dated to the 15th to 12th centuries BCE. Following the Ṛgveda and the Atharvaveda, which consists mainly of hymns to the gods, the first prose texts emerged, focusing on the explanation of ritual actions. These are the Maitrāyaṇī Saṃhitā (MS), the Kāṭhaka-Saṃhitā (KS), and the Taittirīya-Saṃhitā (TS), which belong to what is known as Kṛṣṇa Yajurveda. A text referred to as Kapiṣṭhala-Kaṭha-Saṃhitā is also known, but its content largely coincides with that of the KS. These works are the most important sources for understanding the earliest form of Vedic ritual. The three recensions are the canonical texts of distinct schools (*śākhās*), each associated with a particular brahminical lineage and geographic region. The schools maintained their recensions through strict oral transmission designed to preserve the text including its sound shape.

The contents of the three texts are structured in parallel and share the same topics, namely various Vedic rituals. It has long been assumed that they derived from a common source, branching into an MS–KS group on the one hand and TS on the other, with MS regarded the oldest, followed by KS and then TS (Witzel, 1997).

However, Amano's studies of the MS (Amano, 2014–2015, 2020) have demonstrated that these texts were gradually compiled over a period of approximately one to two centuries. During this time, the relationships among the three schools changed, making it impossible to explain their formation by a simple branching model or by a straightforward relative chronology. The date of composition and the inter-school relationships differ from section to section; thus, inter-school relationships on a section-by-section basis is crucial for clarifying the formation of these texts and furthermore the development of Vedic rituals.

To address this issue, we have undertaken the detection of intertextual similarity. In Miyagawa et al. (2024), we analyzed intertextual similarity between MS and KS in three ritual sections using Word2Vec, the R package `stylo`, and TRACER. The results show that similarity varies by section and suggest that the degree of similarity may provide insights into the inter-school relationships and relative chronology at the time of composition.

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 41–49

2 Aims of the Present Comparison

In this paper, we examine intertextual similarity among the three texts MS, KS, and TS. By extending the analysis beyond a two-text comparison, we expect to gain further insight into the social and historical circumstances of the period. Previous scholarship has long recognized the close relationship between MS and KS, and more recent studies suggest that according to younger sections, KS and TS may in fact show a close relationship, while MS and TS do not show such a relationship (Amano, 2019, 2020). We therefore analyze all three pairwise combinations—MS–KS, MS–TS, and KS–TS.

However, the aim of this study is not merely to measure degrees of similarity, but to investigate how patterns of similarity differ from chapter to chapter. For this purpose, we selected two ritual sections for analysis: Agnyupasthāna (the daily worship of the ritual fires) and Punarādhāna (the re-establishment of the ritual fires). The Punarādhāna sections of MS and KS are known to exhibit particularly strong similarity compared to other sections. It has even been suggested that, rather than representing ordinary parallel transmission, they may reflect a direct borrowing relationship (Amano, 2014–2015). Whether this distinctive relationship can be detected through computational analysis constitutes a major test case for our method.

3 ICoMa: The Analysis Tool

ICoMa (Intertextuality Collation Machine) is a browser-based text reuse detection tool implemented in TypeScript and React. It was designed to address a methodological concern central to this study: different similarity algorithms capture different aspects of textual overlap, and relying on a single metric risks producing method-dependent conclusions. By integrating multiple algorithms in a single platform, ICoMa enables direct comparison of their outputs on identical data, providing a robustness check that has been absent from previous computational studies of Vedic intertextuality.

ICoMa currently supports ten script systems (Coptic, Greek, Latin, Devanāgarī, Syriac, Arabic, Hebrew, Ethiopic, CJK), with automatic script detection and script-appropriate tokenization. The tool is freely available as a web application.

3.1 Tokenization and Word Boundary Handling

ICoMa employs script-dependent tokenization. For alphabetic and abjad scripts (including romanized Sanskrit), word-level tokenization uses Unicode character properties ensuring correct handling of combining diacritics. For CJK scripts, character-level tokenization is applied.

The Sanskrit source texts contain phonological changes known as sandhi, in which, for example, a word-final vowel merges with the initial vowel of the following word. In the plain text, these words remain in their sandhi-merged surface form. Since the present study conducts a character-level analysis, we assume that sandhi does not interfere with the validity of the results and therefore use the plain text with sandhi retained. Word-level segmentation is realized in the lemmatized text, where forms are restored to their dictionary headword entries.

Potential issues concerning word boundaries arise primarily in two cases: whether nominal compounds should be split, and whether verbal prefixes should be separated from their verbs. Nominal compounds were separated as far as possible, while verbal prefixes were generally kept together with the verb. Since such decisions inevitably involve semantic interpretation, it is not possible to apply a single mechanical principle in every case. However, manual adjustments were made to ensure that no differences in segmentation policy arise among the three texts.

3.2 Similarity Algorithms

ICoMa implements six similarity algorithms. All sliding-window algorithms operate on windows of n consecutive tokens (set to $n = 3$ in this study).

Levenshtein distance (Levenshtein, 1966): normalized edit distance between windowed token sequences, yielding a percentage similarity score sensitive to substitutions, insertions, and deletions at the character level.

Character n-gram: token windows are concatenated into a single character string before computing Levenshtein distance. This is designed to mitigate word boundary inconsistencies by operating on continuous character sequences. Under the current parameters ($n = 3$), however, this algorithm produces results identical to the standard Levenshtein distance (see Section 4.4).

Word n-gram: Jaccard coefficient over normalized token sets within sliding windows, capturing

lexical overlap independently of word order.

Jaccard index (Jaccard, 1912): character-level Jaccard similarity over windowed text segments, measuring the ratio of shared to total unique characters.

Smith–Waterman alignment (Smith and Waterman, 1981): local sequence alignment at the token level with fuzzy matching (threshold: 80% character similarity), identifying the single best local alignment between two token sequences.

Script-aware normalization: script-specific preprocessing (Coptic supralinear stroke reduction, Greek polytonic accent normalization, Devanāgarī accent mark and sandhi stripping, and general diacritic normalization) followed by Levenshtein distance. This method is labeled "Coptic-Aware (Specific)" in the ICoMa interface, reflecting its origins in Coptic text reuse analysis (Miyagawa, 2022), but the normalization pipeline generalizes to other scripts.

3.3 Visualization

ICoMa provides six interactive visualization modules: (1) a parallel text viewer with highlighted matches, (2) a heatmap of pairwise token similarity, (3) a dispersion plot, (4) a similarity histogram, (5) a network graph of aligned tokens, and (6) an alignment flow diagram. Results are exportable in JSON format.

The dispersion plots are particularly informative for philological analysis. Each point represents a matched token pair, plotted by position in Witness α (x-axis) and Witness β (y-axis). Dense diagonal clustering indicates that the two texts preserve matching material in the same sequential order; off-diagonal scatter indicates rearrangement or formulaic reuse in non-corresponding positions.

3.4 Data Preparation

The corpus consists of MS, KS, and TS texts in romanized transliteration for the two ritual sections. Two text forms were prepared: a *plain* form preserving the original surface text, and a *lemmatized* form with words reduced to dictionary citation forms. The texts in lemmatized form are provided by Oliver Hellwig, who lemmatized the plain texts using the latest method described in Nehrdich et al. (2024).

3.5 Experimental Setup

All six algorithms were applied to each pairwise combination (MS–KS, MS–TS, KS–TS) for both ritual sections in both text forms. Window size was fixed at $n = 3$. Two metrics are reported: *coverage*, the percentage of tokens participating in at least one match above the similarity threshold; and *mean similarity*, the average score across all detected matches.

Thresholds were calibrated for each tool so that the similarity score for the MS–KS Agnyupasthāna (lemma text) would be close to 50%. The same thresholds were then consistently maintained for all analyses conducted in the same text form within each respective tool. This comparison was expected to fall roughly in the middle among the six pairs, thereby enabling an effective detection of both lower and higher degrees of similarity. As a negative control, a chapter from the MS that differs in both style and content (Darśapūrṇamāsa-Mantras) was included in the analysis.

4 Results

Table 1 presents coverage and mean similarity for each algorithm, text form, and text pair. Character n-gram results are omitted because they are identical to Levenshtein under the current parameters (Section 4.4).

4.1 Cross-Recension Similarity Rankings

All five independent algorithms and both text forms yield the same pairwise ranking without exception: MS–KS > KS–TS > MS–TS. This ranking holds for both Agnyupasthāna and Punarādhāna. The consistency is notable given that the algorithms measure fundamentally different properties—edit distance, character overlap, token-set intersection, local alignment, and diacritic-normalized edit distance—yet they converge on the same conclusion. All the results of the analysis are shown in Table 1.

The negative control confirms that the detected reuse is not attributable to general Sanskrit lexical overlap. Control coverage ranges from 1.9% to 3.1% across algorithms, well below the 47–55% observed for MS–KS Agnyupasthāna. Jaccard is the exception, with control coverage reaching 5.7% (lemmatized) and 13.0% (plain), reflecting its sensitivity to high-frequency character sequences common across any Sanskrit text. This indicates that Jaccard alone is insufficiently discriminating for Vedic text reuse analysis.

The ranking MS–KS > KS–TS > MS–TS is consistent across both coverage and mean similarity. However, it is coverage that highlights the dif-

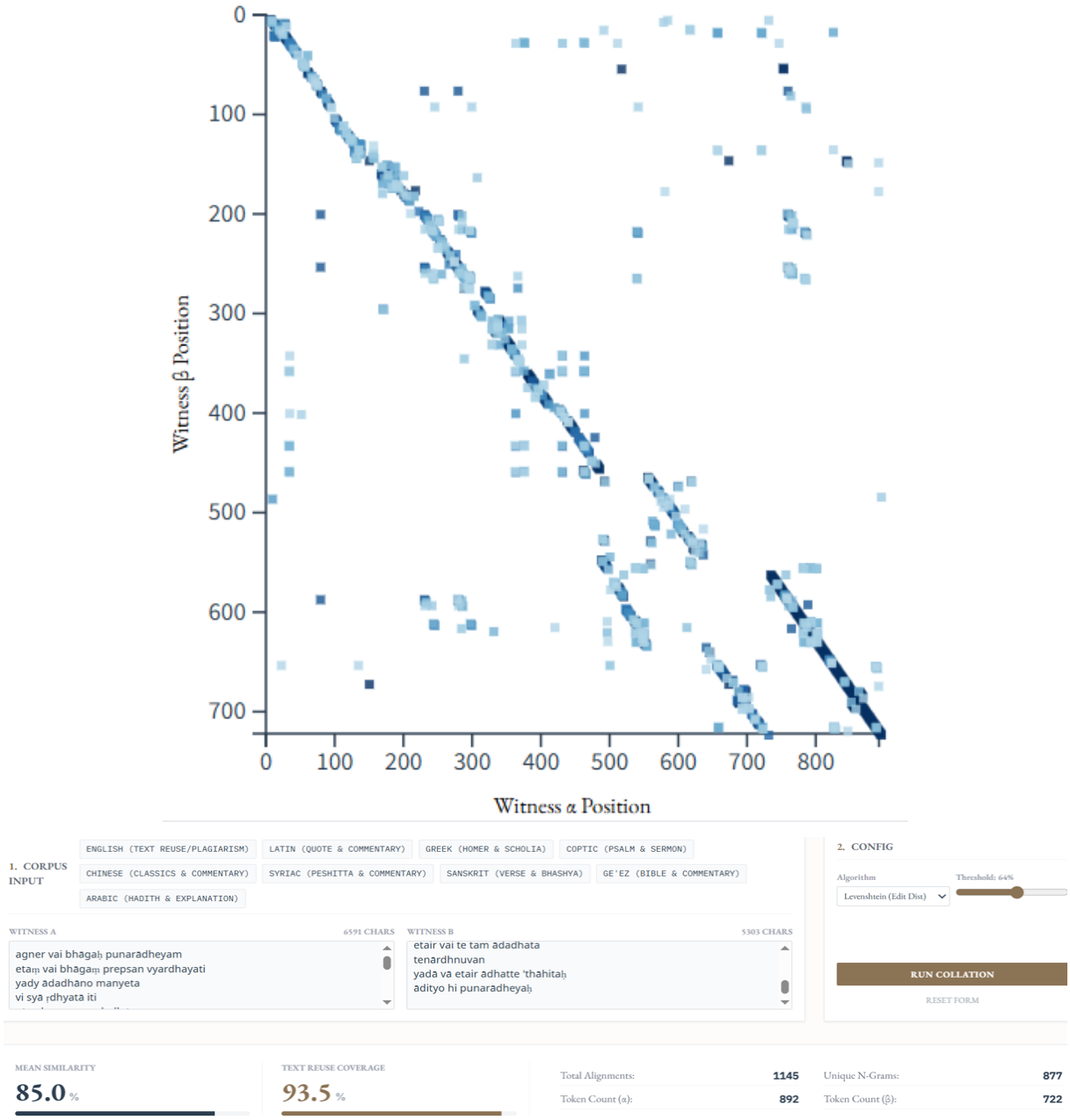

Figure 1: ICoMa interface: collation of MS and KS Punarādhāna (plain) using Levenshtein distance (threshold 64%). Coverage: 93.5%, Mean Similarity 85%, 1145 Alignments. Token counts: 728 (Witness α, MS), 722 (Witness β, KS). The dispersion plot shows dense diagonal clustering, confirming sequential correspondence.

		Reuse Coverage (%)				Mean Similarity (%)				
Algorithm	**Form**	MS–KS	MS–TS	KS–TS	Ctrl	MS–KS	MS–TS	KS–TS	Ctrl	**Thr.**
Levenshtein	lem/agnyup	50.2	30.3	36.3	2.5	91.7	91.4	90.4	91.0	78
	lem/punar	82.9	19.2	20.1	—	96.5	88.6	89.9	—	
Jaccard	lem/agnyup	49.4	31.9	33.4	5.7	99.4	99.4	99.3	99.2	92
	lem/punar	82.3	20.5	22.4	—	99.6	98.8	99.6	—	
Word n-gram	lem/agnyup	47.2	26.6	30.3	2.0	86.7	86.4	87.7	76.5	52
	lem/punar	80.5	18.8	22.2	—	88.4	77.2	76.8	—	
Script-aware	lem/agnyup	50.8	31.3	37.6	3.1	92.2	91.9	91.3	88.7	81
	lem/punar	82.9	20.5	21.0	—	96.8	89.9	91.2	—	
Levenshtein	pln/agnyup	50.1	31.9	36.2	1.9	80.2	77.8	77.5	85.2	64
	pln/punar	93.5	20.5	20.7	—	85.0	76.5	77.5	—	
Jaccard	pln/agnyup	51.0	31.0	33.2	13.0	93.4	93.1	93.4	91.4	86
	pln/punar	87.6	31.4	28.1	—	95.2	90.2	91.0	—	
Word n-gram	pln/agnyup	55.1	31.8	34.2	2.6	54.4	54.3	55.3	60.5	45
	pln/punar	89.6	23.5	25.4	—	59.0	51.6	52.4	—	
Script-aware	pln/agnyup	51.8	33.6	37.5	2.9	81.1	80.0	79.7	90.6	66
	pln/punar	93.4	21.5	23.1	—	87.2	78.1	78.9	—	
Smith-W.	lem/agnyup	3.8	—	—	—	100	—	—	—	20
	pln/agnyup	3.7	—	—	—	100	—	—	—	20

Table 1: Text reuse coverage and mean similarity across algorithms, text forms, and text pairs. "Ctrl" = negative control (MS Agnyupasthāna vs. MS Darśapūrṇamāsa mantra). "Thr." = similarity threshold (%). "lem" = lemmatized, "pln" = plain. Smith–Waterman returns the single best local alignment only. Character n-gram is omitted (identical to Levenshtein; see Section 4.4).

ferences between the pairs more clearly. Because mean similarity calculates the average score across all matches, it is leveled out by the low similarity values found in non-parallel portions of the texts. For this reason, in a study such as the present one—whose purpose is to compare parallel texts—mean similarity is not a particularly effective indicator.

4.2 Contrasting Reuse Profiles of the Two Ritual Sections

The most striking result is the sharp contrast between the two ritual sections (Table 2).

Punarādhāna shows extremely high MS–KS coverage (80.5–93.5%) but low MS–TS and KS–TS coverage (18.8–31.4%). The dispersion plot for MS–KS Punarādhāna (Figure 1) displays dense diagonal clustering, indicating that the two texts preserve matching material in the same sequential order. The small part of offset diagonal lines indicates passages where the order of the sentences has been rearranged, that is, similar descriptions occurring in a different sequence.

Agnyupasthāna shows moderate MS–KS coverage (47.2–55.1%) that is only 15–20 percentage points above MS–TS (26.6–33.6%) and KS–TS (30.3–37.6%). All three pairs share nontrivial textual material, but none approaches the near-total overlap seen in Punarādhāna. The KS–TS comparison (Figure 2) illustrates the corresponding dispersion pattern, with scattered rather than diagonal clustering, reflecting the diffuse, formulaic character of the shared material.

	Agnyupasthāna		Punarādhāna	
Pair	**Lem.**	**Plain**	**Lem.**	**Plain**
MS–KS	50.2%	50.1%	82.9%	93.5%
KS–TS	36.3%	36.2%	20.1%	20.7%
MS–TS	30.3%	31.9%	19.2%	20.5%

Table 2: Contrasting reuse profiles: Levenshtein coverage for the two ritual sections in both lemmatized and plain text forms.

4.3 Suitability of Lemmatized Text and Plain Text for Analysis

If we examine the difference between analyses based on lemmatized text and plain text using coverage as the more reliable indicator, the following result emerges (see Table 2). In the plain text, Punarādhāna MS–KS shows 93.5% compared to 50.1% for Agnyupasthāna MS–KS, while in the lemmatized text the former shows 82.9% compared to 50.2% for the latter (all numerical values based on Levenshtein). This suggests that the difference

Figure 2: ICoMa interface: collation of KS and TS Agnyupasthāna (plain) using Levenshtein distance (threshold 64%). Coverage: 36.2%, Mean Similarity 77.5%, 704 Alignments. Token counts: 1755 (Witness α, KS), 843 (Witness β, TS). The dispersion plot shows that the reuse is scattered throughout the section and is clearly different from the sequential correspondence observed in the MS–KS Punarādhāna (Figure 1).

in similarity is more clearly discernible in the plain text. This tendency can be observed across all tools. The higher similarity values in the plain text indicate a greater degree of surface-level correspondence, that is string identity or phonetic sequence identity between the texts.

Character-level analysis of plain text has the major advantage that the results are not affected by decisions such as how compounds are segmented in the data.

4.4 Algorithm Comparison

The convergence of five independent algorithms on the same pairwise rankings strengthens confidence that the observed patterns reflect genuine textual relationships.

The character n-gram algorithm produced results identical to the Levenshtein distance in every comparison. At the current window size ($n = 3$), concatenating three tokens into a character string before computing edit distance yields the same result as computing edit distance directly. Larger window sizes may break this equivalence and are planned as future work.

Smith–Waterman identifies only the single best aligned region per comparison (3.7–3.8% coverage, 100% similarity), providing limited value for quantifying overall text reuse in its current implementation. Among the sliding-window algorithms, Levenshtein and script-aware normalization offer the best combination of sensitivity and discriminating power.

Jaccard shows one anomaly: its MS–TS coverage rises from 20.5% (lemmatized) to 31.4% (plain), while other algorithms increase by less than 5 points. This parallels its elevated control coverage on plain text (13.0%) and reflects its set-based sensitivity to high-frequency character patterns.

The differences in intertextual similarity are most clearly pronounced by the two tools: Levenshtein shows Punarādhāna MS–KS 93.5% vs. MS–TS 20.5%, and Script-Aware shows Punarādhāna MS–KS 93.4% vs. MS–TS 21.5% (both based on the plain text).

It should also be noted that Miyagawa et al. (2024) analyzed the lemmatized text of the Punarādhāna chapters of MS and KS using a Word2Vec embedding approach, and obtained results that are almost identical to those of the present text-reuse analysis. In the Vedic parallel prose texts examined here, where surface-level correspondences are abundant, character- (i.e., sound-) based similarity analysis can therefore be considered sufficiently effective, to a degree comparable to semantic analysis.

5 Discussion: Implications for Textual History

The contrasting reuse profiles of Agnyupasthāna and Punarādhāna provide quantitative evidence that different ritual sections followed distinct transmission paths within the Kṛṣṇa Yajurveda tradition.

The near-identical Punarādhāna material in MS and KS (coverage up to 93.5%), combined with the sharp divergence from TS (coverage around 20%), suggests that the Maitrāyaṇī and Kāṭhaka schools maintained a particularly close textual lineage for this section. The dense diagonal clustering in the dispersion plot (Figure 1) further indicates that this shared material was preserved not merely as a common vocabulary but in the same sequential arrangement, consistent with the view that the texts had already been fixed and shared almost identical material at the surface-text level, most plausibly through direct borrowing. The TS school appears to have drawn on a different source tradition or substantially redacted the material.

Agnyupasthāna presents a contrasting pattern. The moderate, broadly distributed similarity across all three pairs (MS–KS: 47–55%, KS–TS: 30–38%, MS–TS: 27–34%) suggests that this ritual section circulated more widely among the three schools during an earlier period of shared transmission. The absence of any dramatically close pair implies that all three recensions appear to have shared the same ritual, but because this was a period in which the texts had not yet been fixed, there is relatively little agreement at the surface-text level. The scattered dispersion plot for KS–TS Agnyupasthāna (Figure 2) illustrates this diffuse pattern visually.

The robustness of these patterns across five independent algorithms (Section 4.4) ensures that the conclusions are not artifacts of a particular similarity metric.

6 Conclusion

This study applied multi-algorithm computational text reuse detection to three Kṛṣṇa Yajurveda recensions (MS, KS, TS) across two ritual sections. Three principal findings emerge.

First, all five independent similarity algorithms consistently rank MS–KS as the most closely related pair, followed by KS–TS and then MS–TS, in

both plain and lemmatized text forms. This convergence provides the first multi-algorithm quantitative confirmation of the hypothesis of Amano (2019, 2020) on recension relationships.

Second, it has become clear that, in comparing parallel texts of the Black Yajurveda, a character-level text-reuse analysis based on the plain text is effective. This is because these texts reflect the results of oral transmission and interaction, thereby highlighting a characteristic feature of Vedic literature: the strong emphasis on preserving sound without alteration.

Third, in comparison with other chapters and other intertextual pairings, Punarādhāna MS–KS exhibits a strikingly high degree of similarity and the same sequential arrangement. Unlike the level of parallelism observed in other chapters or combinations, Punarādhāna MS–KS provides strong evidence of the contact status where the texts were fixed, most plausibly a direct borrowing relationship. What had been hypothesized in Amano (2014–2015) can thus be said to have been substantiated by numerical data. In this manner, a way has been opened to explain the differences in similarity as reflecting differences in the periods at which the texts became fixed.

Until now, Vedic parallel texts have generally been understood as branching solely from a common archetype. If the existence of borrowing relationships can indeed be demonstrated, this would offer a new perspective on the formation history of Vedic literature.

The analysis was conducted using ICoMa, a freely available web-based collation tool supporting ten script systems, which is applicable to text reuse research across a range of historical traditions. Future work will implement iterative Smith–Waterman alignment for richer local analysis, and integrate semantic similarity measures through word embeddings to enable detection of paraphrastic reuse. However, the most significant prospect lies in having demonstrated the effectiveness of analysis on plain text, which makes it possible to extend the analysis to other chapters through rapid and streamlined processing with ICoMa. Only when all chapters of the three texts have been analyzed will it become possible to clarify the overall formation process of the Kṛṣṇa Yajurveda.

7 Limitations

ICoMa's word-level tokenization requires consistent word boundary conventions in input data (Section 3.1); the present study controlled for this manually. The negative control was applied only to Agnyupasthāna; extending it to Punarādhāna would strengthen the design. The similarity thresholds were set by manual histogram inspection rather than systematic optimization against gold-standard annotations. The analysis does not distinguish between mantra and brāhmaṇa content types within each section. Finally, the restriction to two ritual sections limits the generalizability of the transmission-historical conclusions.

Ethical Considerations

The texts analyzed are ancient liturgical compositions in the public domain. No human subjects were involved.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP21KK0004.

References

Kyoko Amano. 2014–2015. Zur klärung der sprachschichten in der maitrāyaṇī saṃhitā. *Journal of Indological Studies*, 26/27:1–36.

Kyoko Amano. 2019. The development of the uses of *ha / ha vái / ha sma vái* with or without the narrative perfect and language layers in the old Yajurveda-Saṃhitā texts. *Lingua Posnaniensis*, 61:11–24.

Kyoko Amano. 2020. What is 'knowledge' justifying a ritual action? Uses of *ya evaṃ veda / ya evaṃ vidvān* in the Maitrāyaṇī Saṃhitā. In C. Redard and 1 others, editors, *Aux sources des liturgies indo-iraniennes*, volume 10 of *Collection Religions, Comparatisme – Histoire – Anthropologie*, pages 39–68. Presses Universitaires de Liège, Liège.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

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

So Miyagawa, Yuki Kyogoku, Yuzuki Tsukagoshi, and Kyoko Amano. 2024. Exploring similarity measures and intertextuality in Vedic Sanskrit literature. In

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities (NLP4DH), pages 123–131, Miami, USA. Association for Computational Linguistics.

Sebastian Nehrdich, Oliver Hellwig, and Kurt Keutzer. 2024. One model is all you need: ByT5-Sanskrit, a unified model for Sanskrit NLP tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13742–13751.

Temple F. Smith and Michael S. Waterman. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197.

Michael Witzel. 1997. The development of the Vedic canon and its schools: The social and political milieu. In Michael Witzel, editor, *Inside the Texts, Beyond the Texts: New Approaches to the Study of the Vedas*, pages 257–345. Harvard University.