Recovering the Calcutta Edition of the Mahābhārata for Computational Analysis

Sujoy Sarkar Manoj Balaji Jagadeeshan Pawan Goyal
Indian Institute of Technology Kharagpur, India
sujoysarkarcs@gmail.com, manojbalaji1@gmail.com, pawang.iitk@gmail.com

Abstract

The *Mahābhārata* survives in multiple recensions shaped by a long and complex history of oral and textual transmission. While several major editions, most notably the BORI Critical Edition and Southern recensions,are available in machine-readable form, the Calcutta Edition, the *editio princeps* and principal representative of the Northern recension, remains largely inaccessible for computational research. In this paper, we present a structurally aligned corpus of the Calcutta Edition by reconciling differences between the original printed text and the digitized M. N. Dutta version the book. Using a semi-automated alignment strategy that combines manually verified anchor verses with automated interpolation, we achieve verse-level alignment for approximately 88% of the Calcutta Edition.

Using this aligned corpus, we compare the Calcutta, BORI, and Kumbhakonam recensions through a set of complementary computational analyses. These analyses examine differences in Parva-level structure and expansion, variation in narrative voice based on speaker attribution, and differences in character representations learned from each corpus, together illustrating how recensional variation manifests across structural, narrative, and semantic dimensions. Our findings illustrate both substantial recensional variation in textual volume and structure, as well as relative stability in higher-level semantic patterns. The aligned corpus is released under a Creative Commons license to support future research in Sanskrit philology and computational linguistics.

1 Introduction

The *Mahābhārata*, with its complex history of oral and textual transmission, exists in numerous recensions that have profoundly influenced Indic scholarship. Among the printed iterations of this epic, the *Calcutta Edition* (published 1834–1839) holds a unique status as the *Editio Princeps* (Sukthankar, 1933), the first complete printed edition of the Sanskrit text. Despite its foundational role in early Indology, this edition remains conspicuously absent from the modern digital philological canon. While the BORI Critical Edition (Sukthankar, 1933) and the Southern Kumbhakonam recension (Krishnacharya and Vyāsācārya, 1907) have been successfully digitized for computational analysis, the Calcutta Edition exists primarily as static image scans.

The scale and historical development of the epic highlight the importance of recovering this specific text. As the longest poem in world literature, the *Mahābhārata* comprises approximately one lakh verses, more than seven times the combined length of the *Iliad* and the *Odyssey*, which are traditionally regarded as paradigmatic examples of long ancient epic in the Western literary canon (Khatri, 2023).

The epic text did not emerge as a monolithic entity but evolved through a gradual process of accretion, transforming from a historical narrative into a didactic "reservoir for learning" (Khatri, 2023). Scholarship generally recognizes multiple stages in this textual growth, commonly described as an early core known as the *Jaya*, an expanded narrative referred to as the *Bhārata*,

and a later, more extensive form conventionally designated as the *Mahābhārata* (Khatri, 2023; Buitenen and Fitzgerald, 1973).

This long and fluid transmission history resulted in significant regional variation, broadly categorized as Northern and Southern recensions. Comparative studies report that these variations extend beyond lexical differences to differences in narrative linkage and organization. For example, more than fifty instances have been reported in which narrative transitions present in Southern recensions are absent from corresponding Northern passages (Acharya and Arjuna, 2016).

Within this recensional landscape, the Calcutta Edition stands as the primary printed representative of the Northern vulgate, which dominated 19th-century scholarship and served as the basis for early translations (Krishna Dwaipāyana and Duttā, 1895). While the Critical Edition published by BORI later attempted to distill these variations into a constituted text (Sukthankar, 1933), the Calcutta Edition remains the indispensable historical baseline for understanding the Northern textual tradition.

We aim to address this gap by presenting the first structurally aligned corpus of the Calcutta Edition. At present, the original Calcutta Edition is available primarily as scanned images and lacks a machine-readable text suitable for computational analysis. To enable digital processing, we rely on the electronic text of M. N. Dutta's Mahābhārata (Krishna Dwaipāyana and Duttā, 1895), which was digitized as part of the *Itihāsa* dataset (Aralikatte et al., 2021) and broadly follows the Calcutta recension at the level of narrative content. However, this text differs from the original Calcutta Edition in its organization of volumes, parva divisions, and verse numbering, necessitating systematic structural alignment.

The primary contributions of this work are as follows:

- We present a structurally aligned corpus of the Calcutta Edition of the *Mahābhārata*, obtained by reconciling structural discrepancies between the original printed edition and the M. N. Dutta text through a semi-automated alignment process that combines manually verified anchor correspondences with automatic alignment, resulting in verse-level alignment for approximately 88% of the Calcutta Edition.
- We provide a concise overview of the major digitized *Mahābhārata* corpora currently available for computational research, outlining their recensional affiliations and structural characteristics.
- Using the aligned Calcutta corpus, we conduct a computational comparison with the BORI Critical Edition and the Southern (Kumbhakonam) recension to examine patterns of structural variation.
- We apply Natural Language Processing (NLP) and Machine Learning (ML) techniques, including speaker extraction and embedding-based representations, to analyze structural and semantic variation across these textual traditions.

To facilitate reproducible research, the curated corpus is available at `https://github.com/sujoysarkarai/mahabharatace`.

2 Major Editions and Recensions

The textual study of the *Mahābhārata* has been shaped by numerous printed editions reflecting different recensional traditions and editorial approaches. In the present work, however, we restrict our attention to those editions that are currently available in computer-readable form and are therefore suitable for large-scale computational analysis. These include the BORI Critical Edition [1], representative Southern recensions [2], and the Sastri–Vavilla edition [3], which together

[1] `https://sanskritdocuments.org/mirrors/mahabharata/mahabharata-bori.html`
[2] `https://sanskritdocuments.org/mirrors/mahabharata/mahabharata-sarit.html`
[3] `https://mahabharata.manipal.edu/#/ereader`

provide structured digital access to distinct textual traditions of the epic. While we introduce the Sastri–Vavilla edition here to provide a complete overview of the digital landscape, our downstream comparative analyses focus exclusively on the BORI, Kumbhakonam, and Calcutta editions.

Table 1 summarizes these editions, outlining their recensional affiliations, periods of publication, and digital status. In addition, Table 2 reports the total number of verses in each of these editions, alongside the Calcutta Edition for reference. While many other important printed editions of the *Mahābhārata* exist—including early Northern printings and commentarial versions—most remain available only as physical volumes or scanned copies and fall outside the scope of the present analysis. Notably, the Calcutta Edition, despite its historical importance as the *editio princeps*, is not yet available in a fully machine-readable form, which motivates our effort to recover and structurally align it for computational use.

Edition	**Recension**	**Period**	**Key Contributors & Digital Status**
Calcutta (Editio Princeps)	Northern (Bengal)	1834–1839	**Pub:** Asiatic Society of Bengal. **Note:** First complete printed edition; basis for early translations, including that of M. N. Dutta.
Kumbhakonam	Southern	1906–1914	**Ed.:** T. R. Krishnacharya and T. R. Vyasacharya. **Digital:** Digitized and curated under the SARIT project with XML-based structural markup.
Sastri–Vavilla	Southern	Mid-20th century	**Ed.:** P. P. Sambasiva Shastri. **Pub:** Vavilla Ramaswamy Sastrulu & Sons. **Digital:** The digitization was funded by Vedavyasa Samshodhana Kendra, Subrahmanya.
BORI (Critical Edition)	Critical (Reconstructed)	1927–1966	**Pub:** Bhandarkar Oriental Research Institute, Pune. **Digital:** Keyed and proofread by M. Tokunaga, J. D. Smith, and others; enriched with linguistic annotation.

Table 1: Comparative overview of major *Mahābhārata* editions, their recensional affiliations, and digital availability.

Edition	**Total Verses**
Mahābhārata (BORI Critical Edition)	73K
Mahābhārata (Kumbhakonam Edition)	96K
Mahābhārata (Sastri–Vavilla Edition)	95K
Mahābhārata (Calcutta Edition)	90K

Table 2: Verse counts across selected *Mahābhārata* editions, including digitally available corpora and the Calcutta Edition for reference.

2.1 Textual Structure of the *Mahābhārata*

The *Mahābhārata* is traditionally organized into eighteen major books (*Parvas*), each of which may be further divided into *Upaparvas* and *Adhyāyas*. While this hierarchical structure is broadly shared across recensions, the number and boundaries of Upaparvas and Adhyāyas vary substantially between editions. Such structural variation reflects the epic's complex transmission history and poses a major challenge for direct computational comparison across textual traditions. Table 3 summarizes this variation across the editions considered in this study.

Parva	BORI	Kumbh.	SV.	Calcutta
1. Adi	20 / 225	18 / 260	20 / 216	18 / 234
2. Sabha	9 / 72	8 / 103	8 / 70	10 / 79
3. Vana	16 / 299	16 / 315	17 / 269	21 / 314
4. Virata	4 / 67	5 / 78	4 / 67	4 / 72
5. Udyoga	12 / 197	10 / 196	12 / 186	10 / 197
6. Bhishma	4 / 117	4 / 122	4 / 116	4 / 124
7. Drona	8 / 173	8 / 203	8 / 169	8 / 203
8. Karna	1 / 69	1 / 101	1 / 111	2 / 96
9. Shalya	4 / 64	3 / 66	5 / 59	2 / 65
10. Sauptika	2 / 18	2 / 14	2 / 18	2 / 18
11. Stri	4 / 27	3 / 27	3 / 27	2 / 27
12. Shanti	3 / 353	3 / 375	6 / 338	3 / 367
13. Anushasana	2 / 154	2 / 274	2 / 154	2 / 168
14. Ashvamedhika	1 / 96	3 / 118	3 / 123	2 / 92
15. Ashramavasika	2 / 47	3 / 41	3 / 42	3 / 39
16. Mausala	1 / 9	1 / 9	1 / 8	1 / 8
17. Mahaprasthanika	1 / 3	1 / 3	1 / 3	1 / 3
18. Svargarohana	1 / 5	1 / 6	1 / 5	1 / 6
Total	**95 / 1995**	**92 / 2311**	**97 / 1981**	**98 / 2112**

Table 3: Structural Variation across Editions (Upaparvas / Adhyayas)

2.2 Source Corpus

The *Itihāsa* dataset (Aralikatte et al., 2021) is a large-scale Sanskrit–English parallel corpus containing approximately 93,000 pairs of Sanskrit *ślokas* and their English translations, drawn from the *Rāmāyaṇa* and the *Mahābhārata*. In this study, we use the *Mahābhārata* portion of the dataset, which comprises roughly 73,000 verses and provides a machine-readable Sanskrit text based on the M. N. Dutta edition, broadly following the Calcutta recension. However, differences in structural organization and verse numbering necessitate systematic alignment with the original Calcutta Edition.

3 Methodology

3.1 Manual Structural Alignment

As an initial step, we conducted a manual, chapter-by-chapter inspection of the entire *Itihāsa* Mahābhārata corpus and restructured it to conform to the Parva–Adhyāya organization of the original Calcutta Edition (CE). The M. N. Dutta translation organizes the epic into nine books, which diverges from the eighteen-Parva structure of the CE. Manual inspection revealed several classes of discrepancies, including cases where a single CE Adhyāya was split into multiple units in the digital text, as well as cases where consecutive CE Adhyāyas were merged.

For example, in Book 1, CE Adhyāya 11 was divided into two units in the source dataset, which we corrected during restructuring. Conversely, in Book 6, CE Adhyāyas 23 and 24 were merged into a single unit and were separated to restore the original structure. We also corrected explicit sequencing errors, such as the omission of Adhyāya 257 in CE Volume 12. A complete mapping of all structural adjustments and Parva correspondence is provided in Table 4.

3.2 Manual Verse Number Marking

This study builds upon our previous work (Sarkar et al., 2025), in which we manually aligned verses with Calcutta Edition numbering for the purpose of constructing an entity discovery and

MND Book	MND Adhyāyas	CE Parva	CE Volume	CE Adhyāyas
Book 1	1–234	1. Adi	Vol 1	234
Book 1	1–81	2. Sabha	Vol 2	79*
Book 2	1–315	3. Vana	Vol 3	314*
Book 3	1–72	4. Virata	Vol 4	72
Book 3	1–198	5. Udyoga	Vol 5	197*
Book 4	1–124	6. Bhishma	Vol 6	124
Book 5	1–203	7. Drona	Vol 7	203
Book 6	1–96	8. Karna	Vol 8	96
Book 6	1–65	9. Shalya	Vol 9	66*
Book 6	1–18	10. Sauptika	Vol 10	18
Book 6	1–27	11. Stri	Vol 11	27
Book 7	1–173	12. Shanti (Part I)	Vol 12	173
Book 8	174–365	12. Shanti (Part II)	Vol 12	367*
Book 9	1–168	13. Anushasana	Vol 13	168
Book 9	1–92	14. Ashvamedhika	Vol 14	92
Book 9	1–39	15. Ashramavasika	Vol 15	39
Book 9	1–8	16. Mausala	Vol 16	8
Book 9	1–3	17. Mahaprasthanika	Vol 17	3
Book 9	1–6	18. Svargarohana	Vol 18	6

Table 4: Mapping of M.N. Dutta Books to Calcutta Edition Parvas. *Indicates Parvas where manual Adhyāya restructuring was performed (Splits/Merges).

linking dataset[4]. That effort focused only on the verses containing names listed in book *Index to the Names in the Mahābhārata* (Sørensen, 1904).

As a result, approximately 56,219 verses—corresponding to about 62% of the roughly 90,000 verses of the Calcutta Edition—were manually annotated with Calcutta Edition verse numbers by two annotators. The annotation was carried out by graduate students with school-level Sanskrit education. The annotators were provided with (i) scanned volumes of the printed Calcutta Edition and (ii) the *Itihāsa* corpus, which had been pre-annotated in our earlier work with candidate named entities and tentative Calcutta Edition verse numbers. For each pre-annotated verse, the annotators manually consulted the corresponding passage in the Calcutta Edition, verified whether the verse content matched, and either confirmed the assigned verse number or corrected it where discrepancies were found. This process resulted in a set of manually verified verse alignments that serve as reliable anchors for subsequent automated alignment.

3.3 Automated Interpolation of Verse Numbers

As illustrated in Figure 1, we utilized a simple bounded interpolation method to propagate verse numbers between manually established anchors. For any unmapped segment lying between two anchors (v_{start} and v_{end}), we verify whether the available text lines can be perfectly distributed as standard metrical units.

Specifically, we define the gap in text lines (L_{gap}) and the gap in target verse IDs (V_{gap}) as:

$$L_{\text{gap}} = idx_{\text{end}} - idx_{\text{start}} - 1$$

$$V_{\text{gap}} = v_{\text{end}} - v_{\text{start}} - 1$$

The system automatically assigns sequential verse numbers if and only if the segment satisfies a uniform line-per-verse ratio k:

$$k = \frac{L_{\text{gap}}}{V_{\text{gap}}} \quad \text{where} \quad k \in \{1, 2\}$$

[4] https://github.com/sujoysarkarai/mahanama

MND Volume Number	MND Adhyāya	MND Verse Index	Metrical Line	CE Prava Number	CE Upa-parva Number	CE Verse Number
vol-i	28	0	सौतिरुवाच इत्युक्तो गरुडः सर्पस्ततो मातरमब्रवीत् ।	1	5	1320
vol-i	28	0	गच्छाम्यमृतमाहर्तुं भक्ष्यमिच्छामि वेदितुम् ॥	1	5	1320
vol-i	28	1	विनतोवाच समुद्रकुक्षावेकान्ते निषादालयमुत्तमम् ।	1	5	1321
vol-i	28	1	निषादानां सहस्राणि तान्भुक्त्वाऽमृतमानय ॥	1	5	1321
vol-i	28	2	न च ते ब्राह्मणं हन्तुं कार्या बुद्धिः कथंचन ।	1	5	
vol-i	28	2	अवध्यः सर्वभूतानां ब्राह्मणो ह्यनलोपमः ॥	1	5	
vol-i	28	3	अग्निरों विषं शस्त्रं विप्रो भवति कोपितः ।	1	5	1323
vol-i	28	3	गुरुर्हि सर्वभूतानां ब्राह्मणः परिकीर्तितः ॥	1	5	1323
vol-i	28	3	एवमादिभिरूपैस्तु सतां वै ब्राह्मणो मतः ।	1	5	1323
vol-i	28	4	स ते तात न हन्तव्यः संक्रुद्धेनापि सर्वथा ॥	1	5	
vol-i	28	4	ब्राह्मणानामभिद्रोहो न कर्तव्यः कथंचन ।	1	5	
vol-i	28	5	न ह्येवमग्निर्नादित्यो भस्म कुर्यात्तथाऽनघ ॥	1	5	1325
vol-i	28	5	यथा कुर्यादभिक्रुद्धो ब्राह्मणः संशितव्रतः ।	1	5	1325

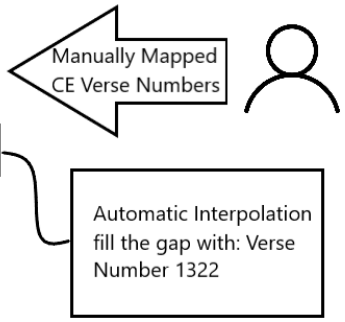

Figure 1: **Semi-Automated Verse Alignment Methodology.** Explicit mappings are established manually for anchor verses (e.g., CE 1321 and 1323), while intermediate gaps are resolved via automatic linear interpolation (assigning CE 1322 to the unmapped segment).

This condition strictly enforces that the gap must consist entirely of standard 2-line *ślokas* ($k = 2$) or, less commonly, single-line units ($k = 1$). Any segment where k is non-integer or deviates from these standard values (indicating complex meters or missing lines) is flagged as `CONFLICT` for manual review.

This automated process aligned an additional 23,406 verses, corresponding to approximately 26% of the corpus. Combined with the manually verified anchors, the dataset achieves verse-level alignment for about 88% of the Calcutta Edition. The remaining 12% of verses have been conservatively left unassigned and flagged for future manual inspection. Table 5 summrizes the result.

Alignment Category	**Verse Count**
Manually Marked Anchors	56,219 (62%)
Automated Gap Filling	23,406 (26%)
Flagged for Future Manual Review	10,376 (12%)
Total Calcutta Edition Verses	**90,001 (100%)**

Table 5: Distribution of verse alignment methods across the restructured Itihāsa dataset.

Evaluation: To evaluate our semi-automated alignment, we audited a random sample of three *Adhyāyas* containing 121 verses (89 manually verified anchors and 32 blind test verses). On the blind test set, the algorithm aligned 30 verses. Manual verification confirmed that 27 were correct, while the remaining 3 instances were assigned wrong verse numbers but were identified as source-text interpolations (extra lines) not present in the Calcutta Edition. The algorithm left 2 verses unassigned due to strict constraints. This yields an algorithmic accuracy of 84.4% and a precision of 100% on valid Calcutta Edition verses (since no valid CE verse was mapped incorrectly). When combining the manually verified anchors with the algorithm's correct predictions, the overall accuracy for the sampled dataset stands at 95.9% (116/121), confirming the high fidelity of the final corpus.

4 Analysis of Textual Discrepancies

During manual inspection and alignment of the corpus, several categories of textual and structural discrepancies were identified. These issues arise primarily from OCR-related errors in the digitized source (Aralikatte et al., 2021), editorial differences in the M. N. Dutta text, and inconsistencies in verse sequencing. Table 6 summarizes the principal types of discrepancies observed, along with representative examples from Volume 1.

Category	Description	Example (Volume 1)
Omissions	Missing names or text segments	Verse 851: the name *Sauti* is omitted; Verse 7019: multiple names (e.g., *Karṇa*, *Śalya*) missing due to OCR failure.
Variations	Spelling or lexical differences	Verse 2634: spelling variation for *Kadrū*; Verse 2533: the name *Mahodara* missing or altered.
Extra Content	Additional text or lines	Verse 4326: extra speaker markers; Verse 130: additional lines concerning *Tumburu* present in the MND text but absent from the Calcutta Edition.
Verse Order	Transposed verse sequences	Verse 8166 appears before Verse 8165 in MN Duttaa.

Table 6: Summary of textual and structural discrepancies identified during corpus inspection.

Beyond these localized discrepancies, we also observed structural irregularities that complicate automated alignment. These include the presence of verses composed in longer meters spanning more than the usual two metrical lines, as well as passages in which verse length alternates irregularly across adjacent sections. Such deviations from the dominant line-to-verse pattern of the *Mahābhārata* introduce ambiguity in automated verse interpolation and necessitate conservative handling during alignment.

Taken together, these discrepancies highlight the challenges inherent in working with large-scale digitized epic texts derived from historical print editions. While many of these issues can be addressed through targeted manual correction or rule-based filtering, others reflect genuine textual variation or metrical diversity and must be accommodated rather than normalized in computational analysis.

5 Comparative Analysis of Recensions

In this section, we investigate structural and narrative differences across three major textual traditions of the *Mahābhārata*: the BORI Critical Edition, the Kumbhakonam (Southern) recension, and the Calcutta recension as represented by the M. N. Dutta (MND) text. While the broad Parva–Upaparva–Adhyāya structure of the epic is shared across recensions, substantial variation exists in the internal organization and textual volume of individual Parvas. The structural differences summarized earlier in Table 3 provide the basis for the comparative analyses presented here.

5.1 Cumulative Verse Growth and Expansion

Verse expansion offers a quantitative perspective on textual variation across recensions. In this analysis, we measure expansion using cumulative counts of metrical lines rather than verse counts. This choice is motivated by the fact that the Calcutta recension, as represented by the M. N. Dutta (MND) text, is not fully annotated with verse numbers, whereas metrical line boundaries are consistently available across all editions. Since a single verse (*śloka*) typically comprises two metrical lines, line counts provide a uniform and comparable proxy for verse-level expansion.

Figure 2 compares the cumulative metrical line counts of the Kumbhakonam and Calcutta (MND) editions against the BORI Critical Edition, which serves as a baseline. The results indicate marked differences in overall textual volume across traditions. The Kumbhakonam recension exhibits the most substantial expansion, with an overall increase of approximately 40% relative to the Critical Edition, whereas the Calcutta (MND) edition shows a comparatively modest expansion of about 6%. At the level of individual Parvas, certain sections display pronounced divergence. In particular, the *Virāṭa Parva*, the *Śānti* and *Anuśāsana* Parvas within the Southern tradition have significant expansions.

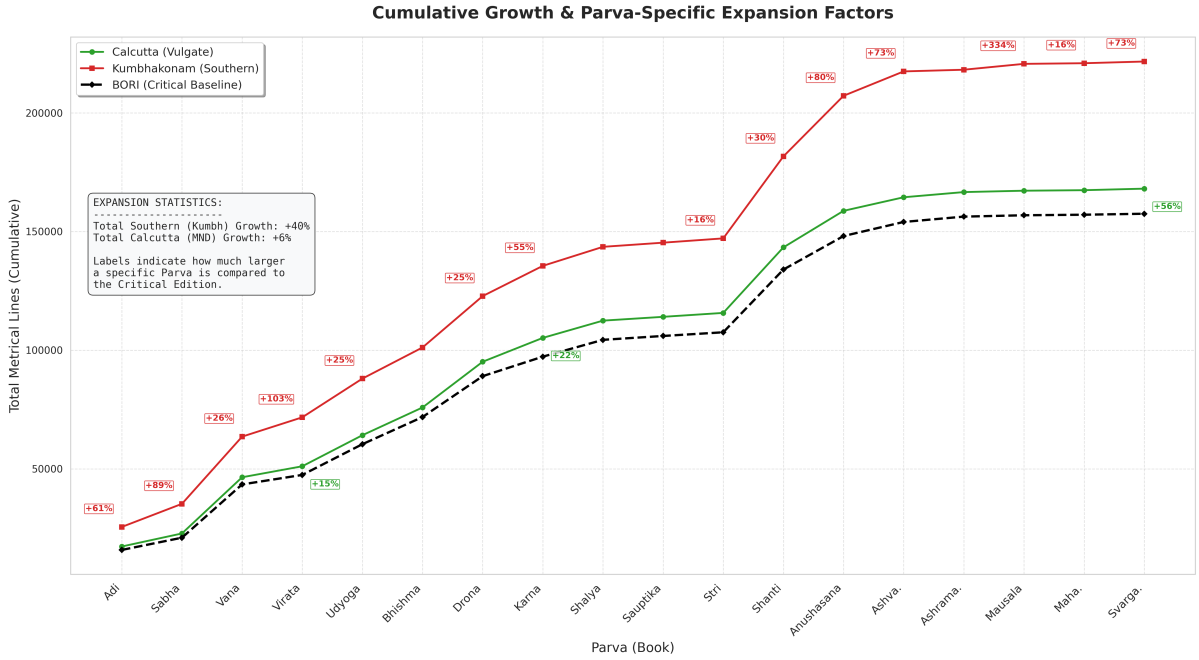

Figure 2: **Cumulative Metrical Lines Growth and Expansion Factors.** The plot compares the cumulative metrical line counts of the Calcutta (MND) and Kumbhakonam editions against the Critical Edition (BORI) baseline. The percentage labels indicate the *Expansion Factor* (growth relative to the Critical Edition) for each individual Parva.

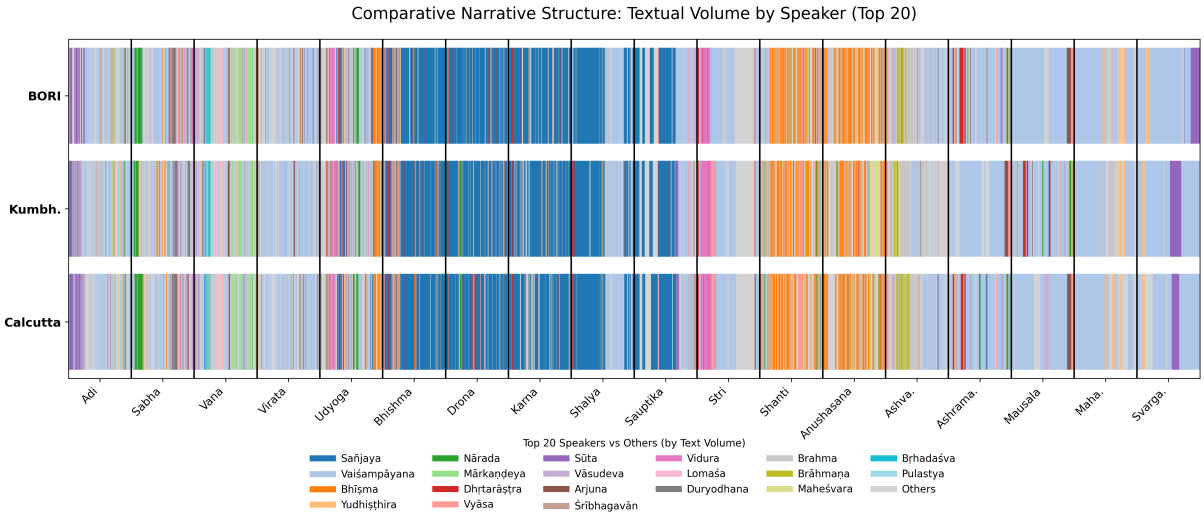

Figure 3: Comparative Narrative Structure: Textual Volume by Speaker (Top 20). The color bands represent the proportional length of speech blocks.

5.2 Speaker-Based Narrative Structure

To examine recensional differences in narrative voice, we analyze the distribution of textual volume across speakers. Speaker attribution is performed using rule-based pattern matching at verse onsets, identifying standard speech markers (e.g., *uvāca*, *ūcuḥ*) and accounting for common sandhi forms. Once a speaker is identified, subsequent verses are assumed to be spoken by the same speaker until a new marker appears. While this heuristic may introduce some noise, it enables consistent large-scale comparison across editions with differing annotation practices.

Figure 3 visualizes the proportional textual volume contributed by the top twenty speakers across the BORI, Calcutta (MND), and Kumbhakonam recensions. The Calcutta Edition largely mirrors the narrative proportions of the BORI baseline, whereas the Kumbhakonam recension exhibits noticeable variations. This visualization offers a coarse-grained view of the overall narrative structure of the *Mahābhārata*, serving as an exploratory analysis; more refined modeling of speaker attribution and discourse structure remains a direction for future work.

5.3 Computational Probing of Character Semantics Across Recensions

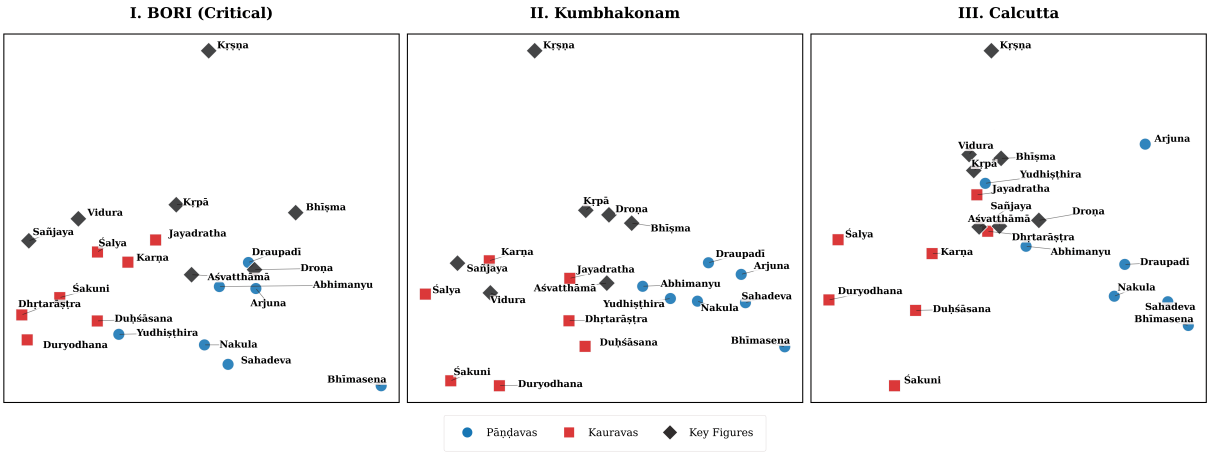

Figure 4: **Character Embedding Space across Recensions.** This projection illustrates the semantic positioning of major characters based on *FastText* embeddings trained independently on each corpus. Clusters represent the *Pāṇḍavas* (blue), *Kauravas* (orange), and Neutral/Others (grey).

To assess whether recensional variation affects computational representations, we conduct an exploratory embedding-based analysis of major characters across editions. This analysis examines whether structural differences between recensions fundamentally alter the underlying semantic relationships and narrative roles of the epic's principal characters.

Separate FastText models (Bojanowski et al., 2017) are trained on the BORI, Kumbhakonam, and Calcutta (MND) corpora, and the relative positions of selected characters are compared. State-of-the-art transformer-based models (Vaswani et al., 2017) are not employed in this analysis for methodological reasons: they produce contextualized embeddings that assign a different vector to each occurrence of a character name, requiring extensive pooling across contexts. In addition, these models are pre-trained on large modern corpora, introducing external semantic biases that are not intrinsic to the epic text. Consequently, static embeddings trained directly on each recension are better suited for the present comparative analysis.

Character embeddings are projected into two dimensions using PCA (Jolliffe, 2011) for visualization (Figure 4). Across all three recensions, the Pāṇḍava and Kaurava groups remain clearly separated, indicating that broad factional structure is consistently encoded across editions. Within this space, Kṛṣṇa appears distant from both groups and approximately equidistant from them, reflecting a distinct semantic profile. Yudhiṣṭhira occupies a more intermediate po-

sition between the two groups compared to other Pāṇḍavas, while secondary figures exhibit greater positional variability across recensions. While exploratory, these observations suggest that recensional variation and narrative structure do not substantially alter the overall semantic space learned by the machine learning models.

6 Conclusion

In this work, we addressed the long-standing absence of the Calcutta Edition of the *Mahābhārata* from computational research by presenting a structurally aligned, machine-readable corpus derived from the digitized M. N. Dutta text. Through a combination of careful manual inspection and conservative semi-automated alignment, we achieved verse-level correspondence for approximately 88% of the Calcutta Edition, while explicitly identifying and isolating segments requiring further manual review.

Beyond corpus alignment, we used this resource to conduct a comparative computational study of three major recensions—the Calcutta, BORI Critical Edition, and Kumbhakonam recension. Our analyses highlight substantial recensional variation in textual volume, Parva-level structure, and narrative expansion, particularly within the Southern tradition. At the same time, exploratory speaker-based and embedding-based analyses suggest a relative stability of higher-level narrative and semantic patterns across editions, despite differences in textual organization.

The aligned Calcutta corpus provides a long-missing digital witness of the Northern tradition and complements existing computational resources based on the Critical and Southern editions. It enables comparative research in Sanskrit philology and computational analysis. By releasing the corpus under an open license, we aim to support broader and reproducible research at the intersection of Sanskrit studies and computational linguistics.

Acknowledgements

This work was supported in part by the National Language Translation Mission (NLTM): Bhashini project of the Government of India. We thank the graduate student annotators for their careful manual verification of verse alignments, which was essential to the construction of the aligned Calcutta Edition corpus.

References

Srinivasa Kumar N. Acharya and S. R. Arjuna. 2016. Grammatical analysis and subject indexing of the mahābhārata and tātparyanirṇaya. In *Proceedings of the Sanskrit Computational Linguistics Workshop (SCLWS)*.

Rahul Aralikatte, Miryam de Lhoneux, Anoop Kunchukuttan, and Anders Søgaard. 2021. Itihasa: A large-scale corpus for Sanskrit to English translation. In Toshiaki Nakazawa, Hideki Nakayama, Isao Goto, Hideya Mino, Chenchen Ding, Raj Dabre, Anoop Kunchukuttan, Shohei Higashiyama, Hiroshi Manabe, Win Pa Pa, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, Katsuhito Sudoh, Sadao Kurohashi, and Pushpak Bhattacharyya, editors, *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 191–197, Online, August. Association for Computational Linguistics.

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.

Johannes Adrianus Bernardus Buitenen and James L Fitzgerald. 1973. *The Mahābhārata*, volume 1. University of Chicago Press.

Ian Jolliffe. 2011. Principal component analysis. In *International Encyclopedia of Statistical Science*, pages 1094–1096. Springer, Berlin, Heidelberg.

Tilak Bahadur Khatri. 2023. Historical development of the epic mahābhārata. *Patan Prospective Journal*, 3(01):173–180.

Krishna Dwaipāyana and Manmatha Nāth Duttā. 1895. *Mahābhārata*. Elysium Press, Calcutta.

Tonape Ramacharya Krishnacharya and Ṭī Ār Vyāsācārya. 1907. *Sriman Mahabharatam: a new edition mainly based on the South Indian texts, with footnotes and readings*, volume 5. "Nirnaya-sagar" Press.

Sujoy Sarkar, Gourav Sarkar, Manoj Balaji Jagadeeshan, Jivnesh Sandhan, Amrith Krishna, and Pawan Goyal. 2025. Mahānāma: A unique testbed for literary entity discovery and linking. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24970–24984, Suzhou, China, November. Association for Computational Linguistics.

Søren Sørensen. 1904. *An Index to the Names in the Mahabharata: With Short Explanations and a Concordance to the Bombay and Calcutta Editions and P.C. Roy's Translation*, volume 1. Williams & Norgate, London.

Vishnu Sitaram Sukthankar. 1933. *Prolegomena [to the critical edition of the Ādiparvan, Book 1 of the Mahābhārata]*. Bhandarkar Oriental Research Institute, Poona.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.