Using Vedalakṣaṇa Granthas for the validation and normalization of Vedic Corpus

Team Svarupa
Soulax Software Pvt.
amruta.barbadikar@soulax.io

Abstract

The Vedalakṣaṇa texts provide lexical, phonological and statistical documentation of the various features of Vedic Sanskrit including sandhi, segmentation, morphological analysis, accent, etc. on the Vedic texts. They provide a standard for measuring the quality of the Vedic texts, especially in the oral rendition of the Vedas, in order to ensure phonetic and structural integrity. In the modern transition to digital formats, maintaining this precision requires more than simple digitization. It necessitates a rigorous validation framework. This paper explores the dual-methodology of manual validation and traditional computational validation using the Vedalakṣaṇa texts, the indigenous rule-based manuals governing Vedic phonology, sandhi, and accentuation.

1 Introduction

The Vedas form the foundational corpus of Indian knowledge traditions and represent some of the earliest preserved textual material in the Sanskrit language. They have been transmitted across generations primarily through oral tradition, supported by well-defined recitational systems that ensured a high degree of precision. Over time, written and printed editions of the Vedas emerged, followed by their digitization in recent decades. In different regions of the country, Vedas are preserved and practiced through multiple śākhās (recension). These traditions exhibit variations in pronunciation, notation, and sometimes different sequence of Mantras and representation, especially when comparing recitation-based versions with analysis-oriented editions. The Vedic literature is a huge collection of thousands of mantras widely divided within 4 Vedas, viz. Ṛgveda, Yajurveda, Sāmaveda and Atharvaveda. The Yajurveda is further divided into two major streams namely, Śuklayajurveda and Kṛṣṇayajurveda. This huge literature has been transmitted across generations primarily through an oral tradition, supported by well-defined recitational systems that ensured a high degree of precision. The Vedamantras occurred in the 'Saṁhitā' form, which means all words and sounds are connected with sandhis and chanted without any break. Although while in the representation we may see the spaces and daṇḍas. For example,

> **Mantra:** a̱gnimī̤ḷe pu̱rohi̇taṃ ya̱jñasya̱ de̱vam ṟ̱tvija̱m | hotā̤raṃ ratna̱dhāta̱mam ||
> (Ṛgveda, 1.1.1)

To understand and preserve the saṁhitās, the ācāryas split them into the 'Padapāṭha' that is the word segmentation marking the word entries, splitting the Sandhi and compound. For example,

> **Padapāṭha:** a̱gnim | ī̤ḷe̱ | pu̱raḥ'hi̇tam | ya̱jñasya̱ | de̱vam | ṟ̱tvija̱m | hotā̤ram | ra̱tna̱'dhāta̱mam || (Ṛgveda, 1.1.1)

The Padapāṭha of the saṁhitā are used to perform eight 'Vikṛtipāṭhas'[1], which is a means of

[1] jaṭā mālā śikhā rekhā dhvajo daṇḍo ratho ghanaḥ |
aṣṭau vikṛtayaḥ proktāḥ kramapūrvā manīṣibhiḥ ||

preservation of Vedas.

Another important instrument of the preservation of the Vedic literature is the Vedalakṣaṇa text. The Lakṣaṇa-texts provides a structured and rule-based way to check the internal consistency of Vedic texts, especially in places where the reading of Saṁhitā becomes unclear or phonetically ambiguous.

A fundamental challenge in the Vedic oral tradition is sandhi ambiguity. In the Saṁhitā texts, mantras are chanted continuously, and adjacent sounds merge according to sandhi rules. This merging creates uncertainty when attempting to reconstruct the original words, particularly in the Padapāṭha. For example, consider the following from Ṛgveda-saṁhitā (8.102.17):

Mantra: taṃ tvā́jananta
Padapāṭha: tam | tvā | ājananta |

From saṁhitā, it is unclear whether the second word is ajananta (with short a), ājananta (with long ā), or another variant entirely. Such ambiguities directly affect meaning, grammatical analysis, and ritual interpretation. 'Āvarṇi' Vedalakṣaṇa addresses this problem by providing systematic methods for preserving the original word forms and their pronunciation. This Lakṣaṇā stores all the words that starts with 'ā', and the word 'ājananta' is also found in this list. Through such detailed rules, padapāṭha segmentation, and cross-references in commentaries and lexical lists, it enables the accurate reconstruction of individual words from continuous chanting. By resolving sandhi ambiguities in a consistent and verifiable way, Vedalakṣaṇa ensures that the oral transmission of the Vedas remains precise, preventing gradual corruption of meaning or pronunciation over generations.

Across different regions of the country, the Vedas continue to be preserved and practised through multiple śākhās and recitational traditions. These traditions show variations in pronunciation, notation, and in certain cases even in textual representation, particularly when recitation-based versions are compared with analysis-oriented editions. While this diversity reflects the richness and continuity of the Vedic tradition, it introduces challenges when attempting to create a uniform digital representation of the texts. Over time, written and printed editions of the Vedas came into existence, and in recent decades these texts have also been digitised.

A clean and reliable Vedic corpus becomes especially important in the context of both traditional scholarship and computational research. Vedic texts involve complex phonetic features such as svara, precise syllabification, and strict mantra boundaries. Errors in any of these components can affect chanting practices, linguistic interpretation, and computational analysis. Hence, maintaining textual accuracy is not only desirable but essential. When building the datasets, it becomes necessary to preserve the most accurate and consistent version of the text, while also clearly identifying and documenting genuine traditional variations. Validation plays a central role in this process, as it aims to maintain the distinction between authentic śākhā-based differences and rectify the errors introduced through digitisation, OCR, or manual transcription.

Svarupa[2] is an initiative towards collecting and curating clean data of the Vedic literature. In this research, we present the manual validation done by Sanskrit scholars with the help of authentic printed sources. The manual validation is backed up by another method of validation traditionally devised to remember and retain the specific cases occurring in the Vedic literature using Veda-Lakṣaṇa text.

[2] `https://svarupa.org/`

2 Literature Survey

Several digital platforms now provide access to Vedic texts, but their quality and reliability vary considerably. Printed editions[3] remain the most trusted sources, yet researchers increasingly rely on digital versions for convenience and large-scale analysis. Because of this shift, it is important to understand what each platform offers and where its limitations lie.

VedaWeb[4] is one of the more academically oriented platforms. It includes selected Vedic texts with information on *pada-pāṭha*, accents, and basic linguistic analysis. However, its coverage is restricted to a few texts and does not include the entire Vedic corpus. The Vedic Heritage Portal[5] hosted by IGNCA brings together texts, audio recordings, and material from different recitation traditions. While this is useful, the portal often provides multiple versions of the same text and the encoding style is not uniform, which creates difficulties when processing the data automatically.

Other widely used repositories are SanskritDocuments[6] and Wikisource.[7] SanskritDocuments is community-driven and therefore the texts lack uniformity; some files are carefully checked, while others contain errors or use inconsistent notation. Wikisource provides Vedic texts in Unicode Devanāgarī, but since the entries are contributed by volunteers, the accent marks, segmentation, and source references are not always consistent. The Digital Corpus of Sanskrit (DCS)[8] is an important resource for Sanskrit studies, but the Vedic material included there is limited and does not contain the detailed svara information needed for precise analysis.

Another online resource hosting Vedic texts is the Vedic Scriptures portal,[9] which presents itself as a platform for reading the Vedas and related materials in digital form. This portal lists the four Vedas and other texts for online access, aiming to make the scriptures available to users via a web interface. While it offers a convenient way to view Vedic content, access may require authentication, and the quality of textual encoding and accuracy of the digitised material varies; the site does not always document source or systematic verification against authoritative editions. Such aspects limit its direct use for computational analysis without further validation and corrections.

These platforms are helpful starting points, but none of them on their own provide a fully reliable and validated Vedic dataset. This makes it necessary to build a dedicated dataset and verify it carefully.

2.1 Collection of data for Svarupa

Each of the existing platforms has inherent limitations. Many sources do not consistently encode Vedic accent marks, or they use non-standard notations. OCR-based data often contains character-level errors, especially in svara symbols and conjunct consonants. In some cases, mantras are incomplete, duplicated, or incorrectly segmented. Due to these issues, it becomes necessary to independently collect the data and validate it against the chosen reference editions. Without systematic validation, the data would not be reliable enough for computational

[3] Printed editions referred to include standard publications such as the editions of Āśvalāyana saṁhitā, Śukla- and Kṛṣṇa-Yajurveda saṁhitās, and other traditionally accepted sources.

[4] `https://vedaweb.uni-koeln.de`. Hosted by the University of Cologne; includes selected Vedic texts with annotations.

[5] `https://vedicheritage.gov.in`. Developed by IGNCA; offers texts, recitations, and multi-śākhā material.

[6] `https://sanskritdocuments.org`. Community-driven repository providing a wide range of Sanskrit texts.

[7] `https://sa.wikisource.org`. A volunteer-supported digital library hosting Sanskrit texts in Unicode.

[8] `https://www.sanskrit-linguistics.org/dcs`. A digital corpus with morphological annotation; Vedic coverage is limited.

[9] `https://vedicscriptures.in`

analysis or scholarly use.

Hence, major data was taken from Vedic Scriptures portal, as the text is accessible by downloading spreadsheets from the portal. Other than that, the collection of Vedic text was carried out using multiple approaches. Web scraping was employed to extract text from publicly available online sources wherever permissible. In cases where digital text was unavailable or unreliable, OCR was performed on scanned printed editions. Additionally, publicly available digital editions of Vedic scriptures were consulted and selectively used.

Since the sources differ in notation, formatting, and completeness, the collected data required further processing before it could be used. The focus was not only on gathering the text, but also on preserving structural information such as mantra boundaries, accent marks, and references.

3 Manual Validation

We normalised the data after collecting it from different sources mentioned before. As a step of preprocessing, the collected data was verified to ensure uniform Unicode encoding, consistent formatting, and clear separation of mantras. This step was necessary to eliminate superficial inconsistencies introduced during web scraping or OCR. Then, the manual validation was carried out. The final dataset taken for validation is as given in Table1.

Table 1: Vedas, Recensions, and Approximate Mantra / Anuvāk Counts

Veda	**Recension (Śākhā)**	**Mantras/Anuvāk**
Ṛgveda	Śākala Śākhā	10,552 Mantras
Sāmaveda	Kauthuma Śākhā	1,875 Mantras
Kṛṣṇa Yajurveda	Taittirīya Śākhā	631 anuvāk
Śukla Yajurveda	Mādhyandina Śākhā	1,975 Mantras
	Kāṇva Śākhā	2,086 Mantras
Atharvaveda	Śaunaka Śākhā	5,987 Mantras

The next step involved finalising an authoritative reference book for each text. All validation decisions were made with respect to these finalised sources, which ensured consistency throughout the process.

The actual validation process consisted of a detailed comparison between the collected data and the reference text. Special emphasis was placed on accent marking, as errors in svara are common in digitized Vedic material. During this process, several types of issues were identified, including wrong syllables caused by typing mistakes or OCR errors, missing words or entire mantras, and extra words or mantras that were inadvertently added. In addition to these, other minor inconsistencies were also observed and corrected wherever possible.

3.1 Ṛgveda

For the Ṛgveda we have a large number of reliable sources available and most of them are consistent and accurate. Due to this availability and consistency, we did not face major difficulties during the validation process. The validation was carried out by considering the VSM (Vaidika Saṃśodhana Maṇḍala, 1933 1951) edition as the primary reference source.

3.2 Sāmaveda

Sāmaveda and Ṛgveda have most of the mantras in common. Therefore, we did not face significant difficulties during the validation process. However, some differences were observed in the svara (musical accent) of the Sāmaveda mantras, which were corrected during validation. The

validation was carried out by the book authored by Dr.Girijaprasad Shadangi (Shadangi, 2024) as the primary reference source.

3.3 Kṛṣṇa Yajurveda (Taittirīya Samhita Textual Variation)

In the Kṛṣṇa- and Śukla-Yajurveda, a special phenomenon of multiple versions of the same saṁhitā are found. One version is based on the representation of the recitation of the mantras. We refer to this as the pronunciation version since it encodes how one should recite it correctly, through various textual cues. For example, the word 'karmaṇe' is represented as 'karmmaṇe'. Here, the doubling of the consonants where the conjunct occurs can be observed. This representation preserves the pronunciation, but for the grammatical analysis and processing, this version needs to be normalised, and we call the normalized format as the analysis version. For the validation, we relied on VSM's Taittirīya saṁhitā (Sontakke and Dharmadhikari, 1970 1972) which is the source for the analysis version.

3.4 Shukla Yajurveda Analysis and Recitation Version Differences

For the Shukla Yajurveda (Mādhyandina and Kānva Shakha), the available dataset is from the analysis version, while the main source book used, which is edited by Svāmī Gaṅgeśvarānanda Udāsīna (Udāsīna, 1994), follows the recitation version. During comparison, differences were observed in consonant usage, similar to Taittiriya saṁhitā, where the recitation version shows doubling of consonants. This mismatch between the analysis-based data and the recitation-based source resulted in validation inconsistencies.

3.5 Atharvaveda (Śaunaka Śākhā)

In the Atharvaveda data, there were not many textual mistakes; however, we faced issues mainly with the svara. For validation we used the book published by Chowkhamba publication (Sātvalekar, 2022) as the source book.

Manual validation of Vedic texts is essential, but it comes with several limitations. The process is slow and requires constant reference to multiple printed editions, which makes it difficult to scale when dealing with large amounts of data. Differences in notation, accent marking, and layout across editions often lead to confusion, and even trained readers may overlook minor variations in svaras or sandhi that are crucial in Vedic material. Because Vedic accents and phonetic rules are highly specific, manual checking alone cannot guarantee uniformity across an entire dataset. This is where Veda-Lakṣaṇa becomes necessary. Its rules provide a stable framework for identifying the correct form of a mantra, including its accents, phonetic structure, and segmentation. Without grounding the validation process in Vedalakṣaṇa, the data may appear complete but still fail to meet the strict standards required for Vedic studies and computational analysis.

4 Veda-lakṣaṇa as a Validation Mechanism for Vedic Texts

Vedalakṣaṇa texts can be utilised for validation mechanisms for the following reasons.

Precise definition Lakṣaṇa texts give clear descriptions of the words, positions, and phonetic environments that produce exceptional behavior. These descriptions offer well defined conditions that can be evaluated computationally.

Direct computational implementation Once these conditions are formalized, any digital dataset (Svarūpa database) can be searched for all matching cases. This produces a complete list of occurrences instead of a selective sample based on printed editions.

Deviations as meaningful signals When the system reports additional or missing cases, each mismatch usually indicates one of the following:

- an oversight in earlier printed editions,

- a Padapāṭha mistake in the digital corpus,
- a different interpretation of segmentation,
- or a genuine ambiguity in the Saṁhitā reading.

These deviations help identify the exact points that need closer examination.

Rules for Complex or Uncertain Cases Vedalakṣaṇa highlights cases where the reading is most likely to be misunderstood. These include final consonant alternatives, uncertain vowel identity, accent doubts and cases where sandhi obscures sound. These are the positions that require validation, and lakṣaṇa supplies the criteria for assessing them.

Cross-Checking with Traditional Texts Each lakṣaṇa rule allows direct comparison between digital evidence and established traditional sources.

Applicability across śākhās and recensions Digital Padapāṭhas generally reflect a single recension, so verification needs criteria that hold across different recitational lines. Lakṣaṇa provides the cross-recensional logic needed for a consistent comparison.

5 Previous Studies on Veda-lakṣaṇa

While there have been numerous efforts towards digitization and qualitative analysis of Vedalakṣaṇa texts, Eichler (2012 2025) is the only digital repository containing the quantitative analysis of the Vedas using the Vedalakṣaṇa. Additionally, (Aithal, 1991) provides a detailed bibliography of all efforts towards Vedalakṣaṇa texts and their analysis. We provide here short descriptions of these two. For convenience, we refer to Eichler (2012 2025) as DE.

5.1 Veda-lakṣaṇa Vedic Ancillary Literature: A Descriptive Bibiliography

In the introduction to the book, Veda-Lakṣaṇa Vedic Ancillary Literature, Sri K. Parameswara Aithal provides an overview of lakṣaṇa granthas, the ancillary texts that define and preserve the characteristics and features of Vedic literature. While many of these works are classified under the Vedāṅga, only a few, such as the śikṣās, belong to its core. Veda-lakṣaṇa texts can be broadly grouped into four categories:

1. works on Vedic phonetics, phonology, and grammar, including prātiśākhya-s and śikṣā-s,
2. anukramaṇi-s and related indices of seers, deities, and meters, sometimes composed in code language,
3. lexical lists documenting words with special grammatical or phonetic characteristics, such as the saptalakṣaṇa, and
4. texts on modified recitation, ranging from manuals for a single recitation mode to comprehensive works describing multiple recitation varieties, like the vikṛti-pāṭhas.

Aithal emphasizes that, despite the availability of manuscripts and printed editions, oral transmission remains the authoritative method, and the systematic rules codified in Veda-lakṣaṇa texts enable accurate memorization and recitation, ensuring the integrity of Vedic knowledge across generations.

5.2 Detlef Eichler's works (DE)

DE's archive (Eichler, 2012 2025) is a foundational resource for the study of Vedalakṣaṇa.[10] The archive places utmost priority on lakṣaṇa treatises, including both widely known and rare, regional texts, particularly from Kerala. These works codify rules for sandhi, recitation variants, vowel hiatus, and other phonetic phenomena, preserving the methods by which Vedic texts have been transmitted orally with precision across generations. By making these texts digitally

[10] `https://sites.google.com/view/vedalakshana/detlef-eichler`

accessible, DE enables modern scholars to study the infrastructure of Vedic oral tradition, resolve ambiguities in recitation, and compare rules across different Vedas and shakhas. The archive includes a wide variety of texts with specialized functions. We utilize DE's analysis as our base for comparison. We apply the conditions proposed by DE for each of the lakṣaṇa texts on the Svarupa dataset and conduct both qualitative and quantitative analysis. Table 5.2 provides the texts preserved in DE. Figure 1 displays the relationship between DE's archive, Vedalakṣaṇa and Vedic oral tradition.

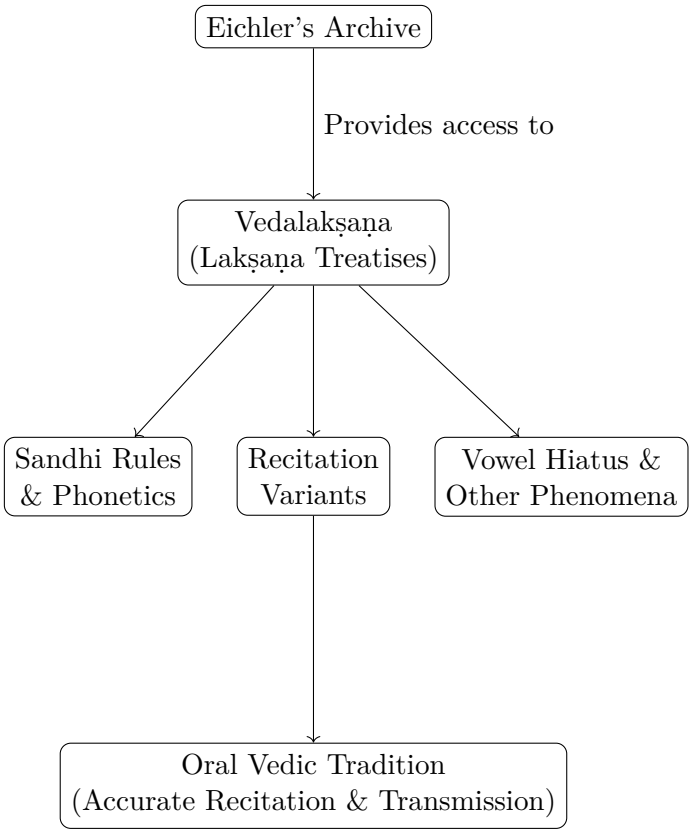


Figure 1: Relationship between Eichler's Archive, Vedalakṣaṇa, and Vedic Oral Tradition

Text / Work	Significance / What It Covers
Avarṇi & Āvarṇi with Commentary	Specialized phonetic/recitation phenomena (irregularities or exceptions) in Vedic recitation.
Itikaraṇa in the Ṛgveda Padapāṭha	Resolving ambiguous readings or variant pronunciations in Padapāṭha recitation.
Vṛthāvyakti with Commentary	Detecting and correcting erroneous or "false expressions" (vṛthā vyakti) in oral transmission.
Śamāna Śikṣā	General rules of pronunciation, phonetics, and recitation for accurate oral transmission.
Sāmavedasaptalakṣaṇa with Commentary	Lakṣaṇa text for Sāmaveda, showing DE's coverage across multiple dimensions.
Antanirdeśa with Commentary (Kṛṣṇa Yajurveda)	Recitation/phonetic treatise for Kṛṣṇa Yajurveda, illustrating lakṣaṇa practices beyond Ṛgveda.
Aruṇaśamāna with Commentary (Kṛṣṇa Yajurveda / Taittirīya)	Śikṣā/lakṣaṇa text for Yajurveda, highlighting recitation norms across Vedic lines.
Kampas in Kṛṣṇa Yajurveda	Addresses "kampas" (recitational/phonetic phenomenon), emphasizing variant recitation modes.
Taittirīya saṁhitā – Hiatus inside Words	Treatise on internal vowel hiatus in continuous chant, crucial for correct pronunciation under sandhi rules.
Trikramalakṣaṇa with Commentary (Yajurveda)	Addresses recitation-mode variation (trikrama, etc.), showing complex recitation traditions beyond the saṁhitā–Padapāṭha dichotomy.
Praṇava Vicāra with Translation & Commentary (Yajurveda)	Focuses on recitation/interpretation of the praṇava (OM), critical for Vedic recitations.
Visarjanīyalakṣaṇa with Commentary (Yajurveda)	Rules for "visarjanīya" — omissions or pronunciation adjustments in recitation.
Svarapañcāśat & Svarasaṁparki with Commentary (Yajurveda)	Deals with svara (tone/intonation) and its recitational relationships, emphasizing pitch and intonation aspects.
Śākhāśamāna with Commentary (Yajurveda)	Lakṣaṇa text addressing shakha differences and standardization across recitation lineages.
Full canonical corpora of other Vedas (e.g., Śukla Yajurveda, Kṛṣṇa Yajurveda) with Śikṣā and lakṣaṇa works	Demonstrates DE's archive spans multiple Vedas, supporting a pan-Vedic understanding of recitation and transmission infrastructure.

Table 2: Vedalakṣaṇa Texts Preserved in Detlef Eichler's Archive

6 Our Approach

The Vedalakṣaṇa texts are majorly available for the Ṛgveda, Sāmaveda, Kṛṣṇayajurveda and Śuklayajurveda. We have considered all the lakṣaṇas provided by DE for the given saṁhitās. We describe ahead the entire procedure to utilize Vedalakṣaṇa texts for our semi-automatic validation tasks. We primarily rely on the padapāṭha without accents for most of the lakṣaṇas,

while for the accent-based lakṣaṇas that deal with accent searches, we utilize the padapaāṭha with accents.

6.1 Rule extraction from Lakṣaṇa texts

The original Lakṣaṇa texts are examined alongside translations and analyses provided by DE, especially the lists of rules, examples, and counts of words. The aim is to make the structure of each Lakṣaṇa rule explicit and suitable for computational use. Each Lakṣaṇa rule is extracted through the following stages.

6.1.1 Analysis

The rule is examined with respect to its scope, stated conditions, exceptions, and underlying assumptions, with attention to whether it relates to phonetic form, accent, or word position. In formulating the rule, DE's interpretations and example-based patterns are followed to understand how the rule has been applied in practice, especially in cases where the original formulation (in the source text) is brief or implicit.

6.1.2 Rule formalisation

Each Lakṣaṇa is formulated as a set of conditions that can be applied to large datasets such as the Svarūpa database. Since the descriptions in the source text and DE's interpretation are generally clear, this step does not involve reinterpretation of the rule, barring a few exceptional cases. Instead, the main task is to convert the rule into a searchable pattern that can be directly applied to the data. For example, the source text states:

> cajayośca thakāraṁ tu hitvā tādi catuṣṭaye |
> makāre ca lakāre ca pare sati vikārabhāk ||
>
> padānto yo makāraśca takāro naparaśca yaḥ |
> dṛśyate yeṣu tau jñātuṁ pravakṣyāmi padānyaham ||

These verses describe cases in which a word ending in 'n' at the end of a 'pada' undergoes modification when followed by specific consonants. To formalise this Lakṣaṇa, we search for all words ending in 'n' in the Ṛgveda Svararahita Padapāṭha. Since the same phonological conditions can also occur within compound words, compounds are included within the search scope. On this basis, the rule is converted into explicit and searchable patterns. The following phonological conditions are defined:

- n followed by c, j, ṭ, d, dh, n, m, l
- n- followed by c, j, ṭ, d, dh, n, m, l.

These patterns allow the Lakṣaṇa to be tested directly against the Svarūpa database. This process ensures that all relevant instances described by the source rule are systematically identified, including cases that may not have been listed in earlier work by DE, as well as repeated occurrences of the same word across different mantras.

6.2 Data preparation and normalisation

Lakṣaṇa rules depend very much on the exact form of the text. Small variations in encoding, word division, or accent marking can change how a rule behaves and may produce misleading results. To avoid this, the textual data is carefully prepared and normalised before any Lakṣaṇa rules are applied. This preparation ensures that the rules interact with the text in a controlled and consistent way, allowing the results to reflect linguistic patterns rather than accidental differences in representation.

6.2.1 Textual alignment across representations

In the Svarūpa database, every mantra is assigned a unique, stable identifier (e.g., 1.1.1). We use this ID to link the Saṁhitā (continuous text) and the Padapāṭha (word-for-word analysis) to the same textual unit. This alignment is critical for validation, since Lakṣaṇa rules are often

defined by Padapāṭha forms but realized phonetically in the Saṁhitā, the shared ID allows us to test conditions across both layers simultaneously. This ensures that every rule is applied to the correct mantra without manual mapping errors.

6.2.2 Textual form and encoding normalisation

The text is standardised to ensure consistency in characters, spellings, and encodings across the corpus. Variations in Unicode representation, Devanāgarī writing forms, and transliteration systems are normalised to support reliable application of Lakṣaṇa rules. This ensures that the rules work correctly and give accurate results, instead of being affected by spelling or encoding differences. Accent symbols are preserved or removed according to the specific requirements of each rule.

The original Sanskrit texts are searched in multiple transliterated forms, including IAST, and WX notation. Each representation is selected based on its suitability for a particular Lakṣaṇa, since different rules place different demands on textual form.

Devanāgarī is preferred in cases where the visual and orthographic structure of the saṁhitā text is crucial. For example, in the padamadhya-vivṛti-lakṣaṇa (hiatus), the task is to identify vowel sequences like a + u which do not undergo further sandhi. Searching in IAST can return incorrect matches as it is not trivial to distinguish this vowel sequence from the vowel au. For example, nama-uktiḥ is present as namauktiḥ with the explicit vowel separation. In such cases, Devanāgarī provides better control over vowel representation and avoids false positives.

IAST is preferred when searching for word endings or phonological environments that depend on segmental structure. For example, to identify words ending in 'a' followed by a specific consonant, a simple pattern such as 'consonant + a' can be expressed compactly in IAST. The same search in Devanāgarī would require an expanded and less manageable character set, making pattern formulation more complex. Therefore, IAST is used where concise and readable pattern definitions are required.

WX notation is used primarily for rules involving accent patterns or large-scale pattern matching, especially in cases related to accent or svara-based Lakṣaṇas. While the same searches could be performed in IAST, WX notation allows stricter and more efficient pattern formulation without ambiguity, making it well suited for computational processing.

Overall, the choice of script or encoding is guided by the nature of the Lakṣaṇa under investigation. Each rule is applied to the textual representation that best captures its linguistic conditions, ensuring both computational efficiency and philological accuracy.

6.3 Database search using extracted rules

After rule extraction and data normalisation, each Lakṣaṇa rule is applied as a structured search over the Svarūpa database. The search evaluates whether the textual data satisfies the conditions defined by the rule, rather than relying on simple string matching.

At this stage, the search is intentionally inclusive. All possible matches are collected so that no valid instance of the Lakṣaṇa is missed. The output consists of a list of mantra identifiers (IDs), the corresponding padapāṭha segments, and the matched words or forms. These results are then used for further mantra-level analysis of the Lakṣaṇa text.

As an example from Napara Vedalakṣaṇa, a source rule is reduced to the following operational patterns:

- n followed by [c | j | t | d | n | m | l]
- n- followed by [c | j | t | d | n | m | l]

While the former is applicable across padas, the latter can be applied only on compound components. For instance, **vṛtrahan'tamam** (Ṛgveda-padapāṭha 1.78.4)

In the case of Napara Lakṣaṇa, the search covers word-final 'n' as well as internal 'n' in compound words when followed by a defined set of consonants. This approach ensures that all potential instances of the Napara Lakṣaṇa are systematically identified for further evaluation.

Thus, the search process involves the following steps:

1. Loading the dataset
2. Applying a regular expression–based search
3. Extracting the matched tokens along with their corresponding mantra identifiers

6.4 Interpretation and refinement of search output

The initial search output may include additional cases beyond those strictly required by the Lakṣaṇa rule. This is expected and is addressed through a controlled refinement process. These steps are applied selectively, as most Lakṣaṇas are resolved through normalisation and rule-based search alone.

6.4.1 Output filtering

The initial results are refined by removing linguistically invalid matches, such as those caused by segmentation errors, editorial symbols, or incompatible contexts. For example, in a Napara Lakṣaṇa search targeting word-final 'n' followed by specific consonants, an automatic search may return cases where 'n' appears at the boundary of an editorial marker or punctuation symbol rather than at a true word boundary. Such matches satisfy the surface search pattern but do not represent valid linguistic instances of the rule. These cases are excluded during filtering.

6.4.2 Adding constraints

If systematic overgeneration is observed, additional constraints are introduced to refine the search results. These constraints are derived either directly from the Lakṣaṇa text or from consistent patterns observed in previously verified examples. This step helps narrow the results without excluding valid cases.

6.4.3 Limited manual verification

For some of the Lakṣaṇas, certain conditions cannot be fully captured through searchable patterns, or specific exceptions are not explicitly enlisted in the earlier steps. In such cases, manual verification is used for final validation. This step is applied only when the number of remaining cases is limited and ambiguity persists. It does not replace rule-based processing but serves as a final check.

In practice, an initial list of candidates is progressively reduced through filtering and constraint application. Only a small number of doubtful cases remain, which are resolved by direct inspection. In some instances, if a rule produces an unusually large number of results, the analysis is further refined by switching to word-based searches within the database. This ensures that both missed cases and overgenerated results are handled in a controlled and transparent manner.

6.5 Validation and correction of results

After refinement, the search results are validated by comparing them with reference sources. This step ensures that the results are reliable and consistent with the earlier work of DE.

6.5.1 Comparison with Detlef Eichler's Corpus

All instances that satisfy the defined Lakṣaṇa conditions are collected during the computational search. In most cases, DE provides comprehensive lists of the relevant occurrences. However, for some of the Lakṣaṇas, DE presents representative examples rather than an explicit enumeration of every occurrence of the same form across the corpus. In such cases, the computational search may identify additional instances that are not explicitly listed in DE but nevertheless conform to the stated Lakṣaṇa conditions. These additional cases are not treated as deviations; instead, they are examined as valid occurrences that naturally arise from an exhaustive corpus-based search. To ensure reliability, these cases are verified against authoritative published sources, including standard critical editions which were previously used during the manual validation stage.

6.5.2 Mismatch analysis

When differences are observed between the search results and DE references, each mismatch is examined in detail. The aim is to determine whether the difference arises from textual variation in the source material, inconsistencies within the dataset, or differences in how the Lakṣaṇa rule has been interpreted or formalised.

6.5.3 Data correction and rerun for final reporting

When issues such as duplicate entries, misaligned identifiers, or missing cases are identified, the dataset is corrected accordingly. Following these corrections, the search is rerun to verify that the updated data yields stable and consistent results.

For example, an initial computational search may produce a higher count than that reported in DE. Detailed examination may reveal that this increase is due to duplicated entries or repeated occurrences of the same lexical item across multiple mantras. After resolving these issues and updating the dataset, the revised count aligns with the validated reference figures. This final rerun serves as a confirmation step and strengthens confidence in both the corrected dataset and the applied Lakṣaṇa rule.

6.6 Discrepancy analysis

During validation, differences were observed between the search results and DE reference listings. These differences arise from a small number of identifiable reasons related to the structure of the data, the search method, and the results reported in DE.

First, some discrepancies are caused by the absence of explicit hemistich boundary markers in the padapāṭha data. A hemistich represents one half of a Vedic verse, and according to Vedic phonological principles, sandhi does not operate between the final word of one hemistich and the initial word of the following hemistich. While hemistich boundaries are indicated in the vedamantra text, they are not currently marked in the padapāṭha.[11] As a result, the algorithm may incorrectly treat words across hemistich boundaries as belonging to a single phonological environment, leading to additional matches that satisfy the formal search pattern but are linguistically invalid. Such cases are identified during validation and excluded from the final count. For instance, in the following example from R.V. 3.31.4, the word **ajānan** occurs at the hemistich boundary and does not undergo sandhi with the next word (tam).

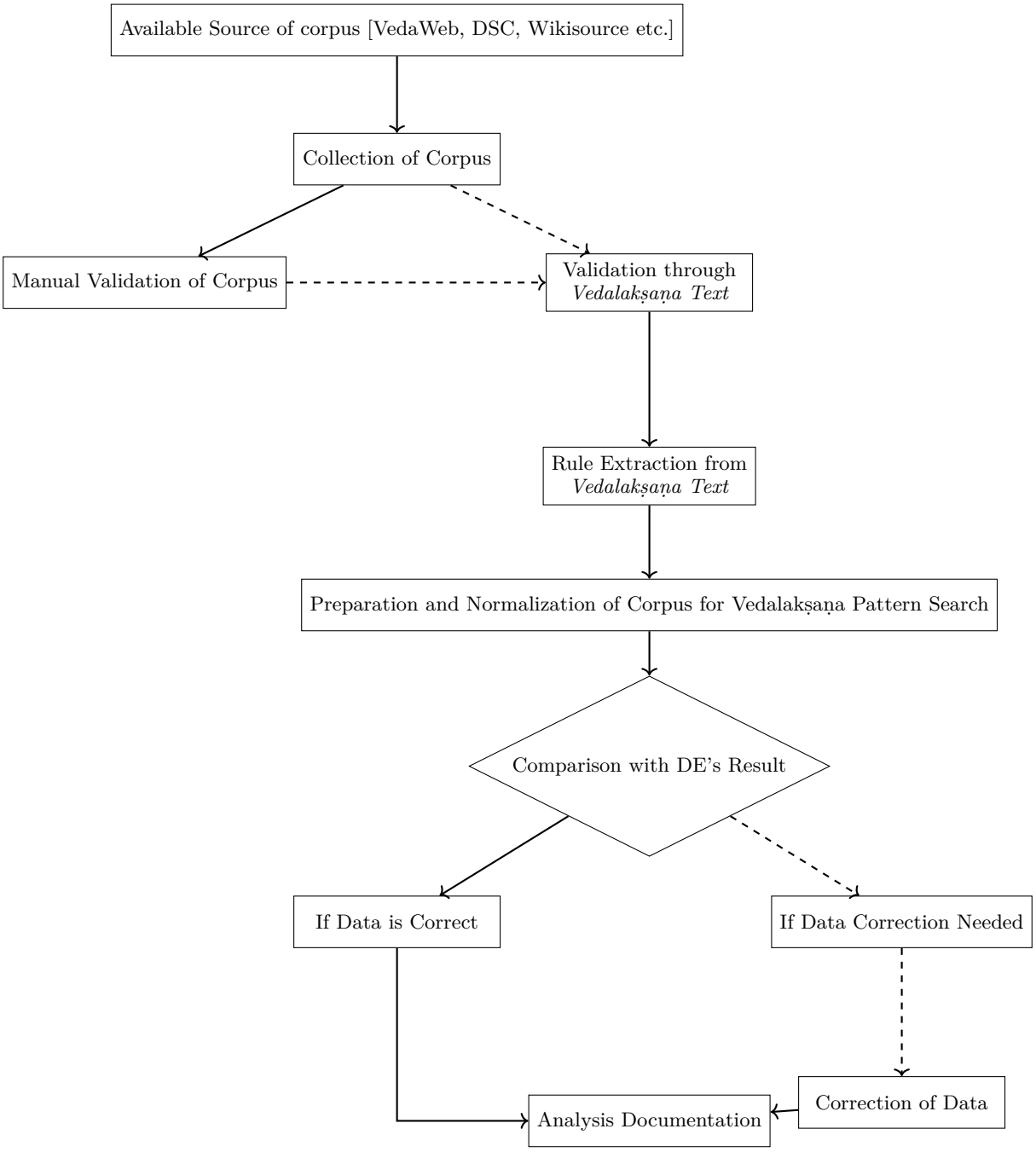
> **Saṃhitā:** jyoti̱stamaso̱ nir**ajānan** | taṃ jāna̱tīḥ
> **Padapāṭha:** niḥ | **a̱jā̱na̱n** | tam

Second, a limited number of discrepancies arise from structural issues within the Svarūpa database, specifically incorrect segmentation or misaligned mantra identifiers. These issues do not involve transcription errors in the text itself but affect how textual units are linked and processed during the search. Such cases are corrected through realignment of segmentation and identifiers, followed by revalidation.

Finally, DE occasionally reports representative examples rather than a complete enumeration of all occurrences. In contrast, the present study aims to document every instance that meets the defined conditions. Consequently, the total number of identified cases may be higher, although the underlying interpretation and application of the Lakṣaṇa rules remain consistent with DE's corpus.

[11] Majority of the books and the digital platforms do not mark the hemistichs of the padapāṭha. And marking it is crucial for such tasks.

Figure 2: Process of Corpus Normalization, Validation, and Analysis



6.7 Illustration: A brief report on the Napara Lakṣaṇa

Definition: Napara answers the question if the word in the Padapāṭha ends in n or m when in the Saṁhitā:

- ñ + (c, j),
- n + (t, d, dh), or
- nasalized l + l is seen.

Search pattern: We use a pattern-based search mechanism where we provide two search inputs, one corresponding to the application of the rules across padas of the padapāṭha (**'n . [c|j|t|d|n|m|l]'**), and the second corresponds to the rule application within a padapāṭha across

compound components (**'n-[c|j|t|d|n|m|l]'**).

Table 3: Pattern searched with example from R.V Padapāṭha for Napara-lakṣaṇa

Pattern	**Saṁhitā-text**	**Padapāṭha-text**	**Example with ID**
n + c	anuyājāmśca kevalānū	anu'yājān | ca | kevalān	**anu'yājān** RV 10.51.8
n + j	aśvinorabudhrañjāmi	aśvinoḥ | abudhran | jāmi	**abudhran** RV 7.72.3
n + n	agmannarastokasya	agman | naraḥ | tokasya	**agman** RV 4.24.3
n + d	akṛṇvandyauṣpitā	akṛṇvan | dyauḥ | pitā	**akṛṇvan** RV 4.1.10
n + ṭ	aryo akṣantā	aryaḥ | akṣan | tāḥ	**akṣan** RV 10.27.8
n + m	satpatīmradabdhānmaho	sat'patīn | adabdhān | mahaḥ	**adabdhān** RV 6.51.4
n-c	vrādhantamo divi	vrādhan'tamaḥ | divi	**vrādhan'tamaḥ** RV 1.150.3
n-m	śatamaśmanmayīnāṃ	śatam | aś-man'mayīnām	**aśman'mayīnām** RV 4.30.20

Comparison and discrepancies: We observe that the number of entries of the napara cases according to DE is 310, and the same for the Svarupa dataset is 400. We discuss the discrepancies ahead.

End of hemistichs: The padapāṭha does not mark the end of a hemistich by default. Search patterns recognized words from different hemistichs as co-occurring together with the possibility of sandhi specified by the napara-lakṣaṇa, resulting in overgeneration. Such cases are manually removed. For example, *akalpayan* R.V.P.P 10.90.1

Errors in Svarupa padapāṭha: There were cases where the conditions overgenerated even after handling the end of hemistich problem. These correspond to errors in the padapāṭha which were corrected. For example, *janitrīn* R.V.P.P 10.35.7

Partial searches: Sometimes subwords and the final component of compounds (like -vān, -van, -san etc) are mentioned by DE instead of complete words. Our search patterns do not consider these as they try to capture the whole pada instead of a partial one. Although, partial pattern searches are possible, they may overgenerate considering all instances of padas ending in the given input. For example, tvā'vān (10.38.5) and saha'vān (1.175.3). These are manually mapped.

7 Directions for Future Work

Having used the Vedalakṣaṇa for the tasks of validation and normalization, we explore possible avenues to which the scope of Vedalakṣaṇa can be expanded. We discuss a few strategies, limitations and utilities of Vedalakṣaṇa texts.

7.1 Veda-Lakṣaṇa for Segmentation Evaluation:

Veda-Lakṣaṇa offers a novel approach to assist segmentation evaluation in Vedic texts by leveraging the prescriptive rules embedded in traditional lakṣaṇa literature. Unlike conventional methods that often split the saṁhitā mechanically or at fixed intervals, this system applies rule-driven analysis to determine the most accurate padapāṭha forms. Each potential segmentation

is assessed against the specific conditions outlined in the lakṣaṇas, allowing the engine to resolve ambiguities and produce a sequence of words that aligns with both phonetic and grammatical conventions of Vedic recitation.

One major limitation is that: all padas of the padapāṭha might not have an associated Vedalakṣaṇa. Although, it is applicable only for a partial evaluation strategy, we see a wider scope of use as the Vedalakṣaṇa texts capture critical instances which are highly ambiguous with respect to sandhi, phonology and morphology.[12]

7.1.1 Why augment Vedalakṣaṇa information on the Padapāṭha?

We discuss ahead what role does the Vedalakṣaṇa information provide in addition to having statistics and heuristics of the padapāṭha, especially for digital platforms like Svarupa, where Vedic resources across various texts and recensions are hosted.

1. **Digital Padapāṭha reflects only one recension:** Most datasets come from a single śākhā, so they show only one segmentation tradition. Anyone working on cross-recensional comparison or looking for alternative analyses needs rule-based segmentation, not a single fixed version. Thus multiple renderings can be realized by multiple sets of Vedalakṣaṇa conditions depending on the tradition.

2. **Digital sources can contain unnoticed human errors:** Typing issues, OCR mistakes, or inconsistent Unicode encoding often enter the data. Apart from a manual validation, the padapāṭhas do not undergo any other check. Thus there is no way to check if a given split is accurate. A rule-driven system can highlight questionable splits and help catch errors.

3. **Padapāṭha shows the result, not the reasoning:** Traditional Padapāṭha does not explain why a split was made. For linguistic analysis or computational models, the explanation matters: which lakṣaṇa rule applied, what phonetic condition triggered it, or whether an exception was involved. While the rules are derivable from the existing differences between the saṃhitā and the padapāṭha, we can also extract ground truth rule applications based on the Vedalakṣaṇa texts.

4. **Veda-Lakṣaṇa helps in resolving ambiguous places, even if it does not validate whole mantras:** Lakṣaṇa texts are not designed to check every mantra end-to-end. But they give precise guidance in places where the oral tradition allows multiple possibilities. When an ambiguous segmentation arises, lakṣaṇa rules let us test which split fits the phonetic and grammatical constraints.

7.2 Need for a New Categorization

Lakṣaṇa literature is extensive, but its rules are spread across many short texts written mainly for practical use by traditional reciters. Modern scholarship often studies these works separately, hiding the deeper linguistic structure shared by the authors. Classifications usually follow textual boundaries or historical order rather than the linguistic nature of the rules themselves.

A linguistically grounded reorganization is helpful because the lakṣaṇas describe different kinds of information: phonetic contrast, accent placement, sandhi patterns, morphophonemic recovery, and even statistical observations about lexical frequency. Grouping them by functional domain rather than by text name reveals the internal logic of the tradition and shows how different lakṣaṇa works complement each other in resolving problems created by the saṁhitā-based recitation, such as vowel contraction, loss of original endings, or accent shifts.

This structure is also useful for computational applications. When rules related to sandhi, accent, or morphological recovery are classified by linguistic type, they can be aligned with tasks

[12] We note and clarify that the rules derived from the Vedalakṣaṇa are not used to perform segmentation, but assist in choosing the intended segmentation given multiple possibilities. It is mainly assisted with the lexical lists that come along with the Vedalakṣaṇa texts, and exclusively useful only for the specific Vedic texts.

such as segmentation, phonological modeling, or tagging of ambiguous forms. A structured view makes it easy to identify which rule sets belong to which level of processing.

8 Conclusion

It is evident that the Vedalakṣaṇa tradition is not merely a historical artifact of the oral tradition but an algorithmic framework capable of solving computational challenges. We have demonstrated that the preservation of Vedic integrity in the digital age requires a synthesis of human expertise and rule-based automation. The work presented here establishes a path forward for the creation of high-quality Vedic corpora through three primary contributions:[13]

Systematic Validation: We have shown that manual validation, while essential, is significantly empowered when paired with the formal constraints of Lakṣaṇa texts. This dual approach effectively rectifies errors introduced by OCR and manual transcription that standard spell-checkers cannot detect.

Linguistic Categorization: By reorganizing the Lakṣaṇa rules into a functional linguistic taxonomy, we have made the traditional wisdom of the Prātiśākhyas and Sūtras accessible for computational modeling, specifically in the areas of phonology and sandhi-splitting.

The Segmentation Evaluator: The development of the Lakṣaṇa-Assisted Segmentation mechanism provides ground truths for rule-driven analyses. While there has not yet been an attempt to use the rules for assisting segmentation, this is a scope for future work.

As we expand the Svarūpa database, the next phase of research will focus on cross-recensional (across śākhā) validation, allowing scholars to document genuine traditional variations while eliminating digital noise.

Acknowledgements

The Svarupa team would like to thank Dr. Bhagyalatha Pataskar, Dr. Jayashree Sathe and Dr. Amba Kulkarni for their invaluable expertise and guidance. Their insights into Vedic grammar, the structural nuances of the Vedas and analysis on the Vedalakṣaṇa mechanisms were instrumental in shaping the direction of this research.

The present work is a part of a larger initiative titled 'Svarupa'.[14] We wish to express our sincere gratitude to Shri Arimilli Ramana Rao for conceptualizing the project and providing the foundational vision that made this study possible.

References

K. V. Abhyankar. 1974. *Veda-Padapāṭha-Carcā: A Study of the Linguistic Analysis of the Padapāṭha.* Bhandarkar Oriental Research Institute, Pune, India. Detailed analysis of Vedic word segmentation.

K. Parameswara Aithal. 1991. *Veda-lakṣaṇa: Vedic Ancillary Literature (A Descriptive Bibliography).* Franz Steiner Verlag, Stuttgart. Also published by Motilal Banarsidass (Delhi, 1993).

Bhagyashree Shreeram Bhagwat. 2021. *Maitrāyaṇī Padapāṭha: A Study.* Vaidika Saṃśodhana Maṇḍala, Pune, India.

R. N. Dandekar and C. G. Kashikar, editors. 1958–1994. *Śrautakośa: Encyclopedia of Vedic Sacrificial Ritual.* Vaidika Saṃśodhana Maṇḍala, Pune, India. A multi-volume comprehensive record of śrauta rituals.

Madhav M. Deshpande. 1975. Critical studies in indian grammarians i: The theory of homogeneity [savarnya]. *The University of Michigan*, 2:1–103.

T. N. Dharmadhikari. 1989. *Yajñāyudhāni: An Album of Implements Used in Vedic Rituals.* Vaidika Saṃśodhana Maṇḍala, Pune, India. Visual and descriptive catalog of ritual tools.

[13] We provide the results of all the analysis at Svarupa-Vedalakṣaṇa, and the validated data is presented on the Svarupa platform.

[14] `https://svarupa.org/`

T. N. Dharmadhikari. 2008. *The Maitrāyaṇī Saṃhitā: Its Ritual and Language*. Adarsha Sanskrit Shodha Samstha, Pune, India. Critical study of the Maitrāyaṇī recension of the Yajurveda.

Detlef Eichler, 2012–2025. *Veda-lakṣaṇa: Traditional Mechanism in Vedic Texts*.

Oliver Hellwig, 2010–2024. *The Digital Corpus of Sanskrit (DCS)*.

Kātyāyana. 1967. *Vājasaneyi Prātiśākhya*. Chowkhamba Sanskrit Series Office, Varanasi. Ancillary text for the Shukla Yajurveda.

S. D. Khadilkar. 2003. *Kātyāyana Śulba Sūtra: With English Translation, Explanatory Notes, Diagrams and Articles*. Vaidika Saṃśodhana Maṇḍala, Pune, India. Crucial for ancient Vedic geometry.

Sriram Krishnan, Sepuri Gayathri, and Amba Kulkarni. 2025. Challenges in processing Vedic Sanskrit: Towards creating a normalized dataset for the Ṛgveda-saṃhitā. In Amba Kulkarni and Oliver Hellwig, editors, *Computational Sanskrit and Digital Humanities - World Sanskrit Conference 2025*, pages 131–147, Kathmandu, Nepal, June. Association for Computational Linguistics.

śrīpād Dāmodar Sātvalekar. 1972. *Shukla Yajurveda: Kanva Saṃhitā*. Chowkhamba Sanskrit Series Office, Varanasi, India, 1 edition. Critical edition with commentary (Vajasaneyi–Kanva recension).

Śrīpād Dāmodar Sātvalekar. 2022. *Atharvaveda Saṃhitā*. Chaukhamba Sanskrit Pratishthan, Varanasi, India. Standard Sanskrit text edition of the Śaunaka recension.

śrīpād Dāmodar Sātvalekar. n.d. *Atharvaveda Saṃhitā*. Nag Publishers, Delhi, India. Sanskrit text edition of Atharvaveda Saṃhita.

Śaunaka. 1927. *Rigveda Prātiśākhya*. Asiatic Society of Bengal, Calcutta. A foundational Vedalakṣaṇa text for the Rigveda.

Girijaprasad Shadangi. 2016. *Samaveda-Padapāṭha*. Dr. Girijaprasad Shadangi, Tirupati, India. First edition research manual for Samaveda word segmentation.

Girijaprasad Shadangi. 2020. *Prakrutigana of Samaveda*. Girija Prasad Shadangi, Tirupati, India. Essential for musical notation validation.

Girijaprasad Shadangi. 2024. *Sāmavedasaṃhitā: Kauthumaśākhīyam (Volume 23 of Series)*. Dr. Girijaprasad Shadangi, Tirupati, India. Primary source for Samaveda validation in modern research.

Girijaprasad Shadangi. n.d. *Ṛktantram: Sāmavedīya Prātiśākhyam (Sabhāṣyam)*. Girija Prasad Shadangi, Tirupati, India. A Pratisakhya of Samaveda with Sanskrit and Hindi commentary.

N. S. Sontakke and T. N. Dharmadhikari. 1970–1972. *Taittirīya Saṃhitā: With Padapāṭha and the Commentaries of Bhaṭṭa Bhāskara Miśra and Sāyaṇācārya*. Vaidika Saṃśodhana Maṇḍala, Pune, India.

Gaṅgeśvarānanda Udāsīna. 1994. *Yajurvedasaṃhitā*. Yogesvara-guru-gangeshava charitable trust, Bombay, India.

Vaidika Saṃśodhana Maṇḍala. 1933–1951. *Ṛgveda-Saṃhitā: With the Commentary of Sāyaṇācārya (Critical Edition in 5 Volumes)*. Vaidika Saṃśodhana Maṇḍala, Pune, India. The standard "Poona Edition" of the Rigveda.

Albrecht Weber, editor. 1972. *Śukla Yajurveda: Madhyandina Saṃhitā*. Chowkhamba Sanskrit Series Office, Varanasi, India. Edition in both Madhyandina and Kanva recensions.

William Dwight Whitney. 1871. *The Taittirīya Prātiśākhya: With its Commentary, the Tribhāṣyaratna*. American Oriental Society, New Haven. Crucial manual for the Krishna Yajurveda.