Challenges in Processing Vedic Sanskrit: Towards creating a normalized dataset for the *Ṛgveda-saṃhitā*

Sriram Krishnan[1], Gayathri Sepuri[2] and Amba Kulkarni[1]
[1] Department of Sanskrit Studies, University of Hyderabad
[2] Svarupa
sriramk8@gmail.com, sepurigayathri16@gmail.com, apksh@uoh.nic.in

Abstract

Computational processing of Vedic Sanskrit is relatively less explored because of various nuances that did not carry forward to Classical Sanskrit. There are many platforms that host source texts of the Vedas along with annotations, but require validation from various aspects. In this work, we present an alignment between the *Ṛgveda-pada-pāṭha* annotations of three platforms viz. Vedic Scriptures, Vedaweb and the Digital Corpus of Sanskrit in order to produce a unified database encompassing the information from all the three platforms. In this process, we observe the challenges in processing the *Ṛgveda-pada-pāṭha*. We also propose a Vedic morphological analysis engine that handles the intricacies of the *pada-pāṭha*, uses the services of the morphological analyzers from Sanskrit Heritage Platform and Saṃsādhanī, along with the annotations of DCS and Vedaweb.

1 Introduction

Sanskrit literature has been classified into two categories: Vedic and Classical. Vedas and their ancillary texts are attributed to the Vedic Period. Grammarians and linguists of the later period prepared treatises like *prātiśākhyas, śikṣā-granthas*, etc. to understand the Vedas. The four Vedas viz. *ṛg, yajus, sāma, atharva* each contain four categories: *saṃhitā, brāhmaṇa, āraṇyaka and upaniṣad.* While historians and linguists differ very much in the time period of each of the sub-categories, from a computational perspective, these can be observed under one banner - Vedic Sanskrit.

Since the Vedas had been distributed across the country and across various cultures, there are various branches or recensions, called *śākhās*. Each of the four Vedas have multiple *śākhās* and according to *Mahābhāṣya*, there were more than 1000 *shākhās* put together. In the present day, though, only a few of them are accessible.

As most of the transition of knowledge was done orally in the ancient times, eleven different metrics were introduced to preserve these texts from tampering or changes. Three are termed *prakṛti-pāṭha* and the remaining eight are referred to as *vikṛti-pāṭha*. These metrics have been operated on the *saṃhitā* of each of the Vedas. The three *prakṛti-pāṭhas* are *saṃhitā-pāṭha, pada-pāṭha* and *krama-pāṭha*. The *saṃhitā-pāṭha* is written in a continuous form where sandhi happens across words.[1] The *pada-pāṭha* consists of individual words of a sentence/mantra along with a few indicators for compounds, prefixes, suffixes, etc. But these *pāṭhas* are available only for the *saṃhitā* category of the Vedas. For the remaining categories (*brāhmaṇa, āraṇyaka* and *upaniṣad*), they have to be obtained either mechanically or computationally.

There are many differences between Vedic Sanskrit and Classical Sanskrit, at various levels in phonology, accent (*svara*), grammar, vocabulary and usage. The most important feature of the Vedic Sanskrit is the accent. Most of the texts in Classical Sanskrit do not retain or prescribe the

[1]Sandhi is a phenomenon of euphonic transformation at the word boundaries.

accents while almost all the Vedic texts are with the accents. From the perspective of grammar, Vedic Sanskrit includes the injunctive and subjunctive moods (*leṭ lakāra*) which are lost in Classical Sanskrit. Multiple infinitives were found in the Vedic literature while the Classical literature has only one. And there are semantic differences in the three synthetic past tenses (imperfect, perfect and aorist). With respect to vocabulary and usage, a lot of words have been introduced after the Vedic period and some words have been lost during this transition. And the hiatus (break between consecutive vowels) was allowed during sandhi and compound formations and also in the interior of words.[2].

Recent efforts to develop Sanskrit processing tools have focused more on Classical Sanskrit, as there are far more intricacies to handle when considering the Vedic Sanskrit. There are various nuances which did not carry forward from Vedic to Classical, in both the grammar as in vocabulary and constructions, and also in usage. The present work is to investigate these nuances from a computational perspective towards analysing how best the existing resources and tools of morphological analysis fare on these texts. We take up the *śākala-śākhā-saṃhitā* of the *Ṛgveda*, and provide a comparison of the *pada-pāṭha* and the morphological annotations from various resources and tools. Majorly, the *Ṛgveda* dataset of the Digital Corpus of Sanskrit (DCS) (Hellwig, 2010 2021),[3] the *Ṛgveda* annotations of the Vedaweb platform (VW),[4] morphological analyses obtained from the Sanskrit Heritage Engine (SH)[5] and Saṃsādhanī platform (SCL),[6] are taken into consideration for the comparison.

In this process, we analyse and observe the limitations of these resources and tools and propose a new database comprising of information from all of these tools and resources. The resultant database contains *saṃhitā-pāṭha, pada-pāṭha*, with and without accent marks, morphological analysis from DCS, Vedaweb, SH, SCL.

Section 2 gives an overview about the existing Vedic resources and tools available for computational processing, along with their limitations. Towards the end it hints at how we are intending to tackle these limitations. Section 3 elaborates about the *Ṛgveda-saṃhitā-pāṭha* and *pada-pāṭha*. Here the datasets, annotations and the differences between the systems (Vedic Scriptures, Vedaweb and DCS) are discussed followed by the details on the alignment between these systems. Section 4 deals wth the challenges present in the *pada-pāṭha*, focusing on *Ṛgveda*, when considered for computational processing. Section 5 provides the architecture details of our Vedic morphological analysis engine and the details of the *Ṛgveda* database constructed from the alignment and the engine. The last section concludes by providing inferences and future directions.

2 Existing Data and tools

2.1 Vedic data

There has been a surge in the digitization of Vedic texts and various platforms have e-books and e-readers of the Vedas. But the computationally accessible (machine-readable) data is quite less and annotations are much lesser. Platforms like Gretil[7] have the original texts along with their commentaries. Platforms like Vedic Heritage Portal[8] provide access to various texts ascribed to the Vedas along with their audio rendering. There are other portals like Vedic Scriptures[9] which provide access to different renditions of the Vedas.

Annotated Data: The Digital Corpus of Sanskrit (DCS) records the lexical, morphological and sentential annotations of the mantras from *Ṛgveda* and *Atharvaveda-saṃhitā*. Vedaweb

[2] *dūraādiśam* → *dūre-ādiśam*, *vasyaiṣṭaye* → *vasyaḥ-iṣṭaye*

[3] http://www.sanskrit-linguistics.org/dcs/

[4] https://vedaweb.uni-koeln.de/

[5] https://sanskrit.inria.fr

[6] https://sanskrit.uohyd.ac.in/scl

[7] https://gretil.sub.uni-goettingen.de/

[8] https://vedicheritage.gov.in/

[9] http://vedicscripture.com/

provides a lot more details such as the morphological analysis, translation, *chandas* (Vedic metres) information, links to lexicon, etc. for the *Ṛgveda-saṃhitā*.

Limitations: While the images of the Vedas and their ancillary texts are available in these platforms, the machine-readable e-texts are very less, and so are the annotations. The existing machine readable annotated data, like the DCS, requires validation and normalization at various levels (Krishnan et al., 2023). With various recensions (*śākhās*) and an innumerable number of editions of each of the Vedas, a validated text which is computationally accessible in one of the existing encoding schemes is a primary concern, followed by lexical, morphological, syntactic and semantic annotations.

This presents to us a necessity for an exhaustive analysis of each of these texts to produce a normalized version which has both the original text and its annotations. The task would be to unify the details presented in each of the platforms, validate and normalize them, and if possible provide new annotations for those which are not yet annotated.

2.2 Resources and Tools

There are various tools for processing Sanskrit texts that help in NLP tasks such as segmentation, morphological analysis, parsing, word-generation, etc. Both grammar-based and machine learning approaches have been incorporated to develop these tools and in recent times, a hybrid of these two have also been developed. Sanskrit Heritage Platform, Saṃsādhanī, Dharmamitra,[10] etc. are some of the tools used for many of the NLP tasks.

Limitations: All these tools have been built considering the texts predominantly from Classical Sanskrit. SH and SCL do not produce results of some of the Vedic-specific constructions. For example, the subjunctives, injunctives and infinitives other than *tumun* are some of the constructions absent in SH and SCL. It is trivial to update the lexicon with stems and basic constructions, but for special constructions, it becomes complicated for systems like SH.

In addition to all these, the representations of these tools vary based on their own design decisions. Saṃsādhanī's analyser produces an output based on a format that is understandable by traditional grammarians. DCS and Vedaweb follow an approach where the annotations use the western linguistic terminologies. While SH retains these terminologies, their representations are based on the traditional grammarians. DCS and Vedaweb have static data, SH and SCL generate the analysis based on their own lexicons. The analysis of SH and SCL have a few other features which are unavailable in DCS and Vedaweb. The differences between the representations of DCS and SH exist at various levels in chunk, word-form, stem, morphological analysis and compounds (Krishnan et al., 2023). The differences in morphological analysis attribute to the absence of secondary conjugations, *gaṇa* (class) information and *pada* (*ātmane / parasmai*) information in DCS, which are present in SH analyses. Differences in tense-mood combinations (*lakāra*) have to be mapped between DCS and SH. SCL produces the analysis similar to SH but uses the traditional linguistic terminology rather than the Tense-Mood representations of both SH or DCS. Thus a mapping between the representations of SCL and SH are to be generated. In the latest DCS representations, the Aorist and Perfect have been clubbed together and represented as Past, leading to multiple possibilities. Many such one-to-many mappings from DCS to SH have to be considered while building the conversion from one system to the other.

The tools which use data-driven approaches require additional validation as some of their predictions might result in incorrect analysis. As the final annotated corpus requires the exact analysis, these methods, when used alone, are not reliable enough to build a fool-proof dataset.

2.3 Our approach

The *pada-pāṭha* serves as the segmentation of the *saṃhitā-pāṭha* along with various phonological and morphological indicators. Any analysis on the Vedic *mantras*, thus starts with its *pada-pāṭha* providing the words and the indicators there in.

[10] `https://dharmamitra.org/`

The Vedaweb *Ṛgveda* dataset has provided various annotations of the *Ṛgveda mantras* and the words (*padas*) extracted from the *pada-pāṭha*. On the other hand, the DCS *Ṛgveda* dataset has provided the lexical and morphological annotations from a sentential perspective. The data from Vedic Scriptures contains the *saṃhitā-pāṭha* and the *pada-pāṭha* along with various metadata about the *mantras*. This calls for an alignment between the Vedic scriptures *pada-pāṭha* and the *pada-pāṭha* proposed by VW along with an alignment of their words with the DCS segments. We thus attempt at an alignment between the Vedic Scriptures data and the annotations of DCS and VW. Due to the design differences between the three systems, the alignment poses various challenges which are discussed in detail in this paper. The aligned dataset contains information from all the three platforms.

In addition to this, considering the various differences due to the design decisions of DCS and SH, an alignment is attempted between the DCS' annotations of the *Ṛgveda* mantras with the SH and SCL morphological analysis of the *pada-pāṭha*. For this, the words have to be extracted from the *pada-pāṭha* entries which requires us to understand the various features of the *pada-pāṭha*. The aligned dataset is extended further to include the possible analyses of these words proposed by SH and SCL, to produce a unified dataset encompassing the features from all the four platforms.

3 Aligning Ṛgveda-saṃhitā and pada-pāṭha with annotations from various resources

3.1 Ṛgveda-saṃhitā and pada-pāṭha

The *Ṛgveda-saṃhitā* consists of ten *maṇḍalas* with 1,028 *sūktas* and 10,527 *mantras*.[11] The traditional methodology of preserving the original versions of the *saṃhitā* involve various representations and algorithms developed by the Vedic schools, broadly categorized under two divisions: *Prakṛti-pāṭha* and *Vikṛti-pāṭha*. *Prakṛti-pāṭha* deploys three representations:

1. The original *saṃhitā* form,
2. *pada-pāṭha* - the individual words of the *saṃhitā-mantras* are represented separately without any occurrence of sandhi with adjacent words, along with additional indicators like compound markers,
3. *krama-pāṭha* - bigrams of the individual words are represented separately where the sandhi occurs within the two consecutive words taken into account for the bigram, but not with their adjacent bigram.

Although the *pada-pāṭha* encompasses all the words of the mantras separately, it cannot be considered as the segmentation of the mantras, as it also encodes information providing clues for disambiguation at word-level, lexicon-level and morphology-level. The *pada-pāṭha's* motive was not only the preservation of the *mantras* intact, but also to analyse the *saṃhitā* from a grammatical point of view. There are several observations when comparing the *pada-pāṭha* with the *saṃhitā-pāṭha* (Pillai, 1941):

1. Resolving sandhi
2. Restoring the original accents of the words: sandhi introduces transformation of the word boundaries which also affect the original accents of the words. The *pada-pāṭha* helps in preserving the words with their original accents.
3. There are instances where the *saṃhitā* transforms certain characters due to special *sandhi* found only in Vedic Sanskrit. The *pada-pāṭha* contains the original version without the transformation. For example *s* to *ṣ* and *n* to *ṇ*.
    - *ūti ṣa bṛhato divo* → *ūti / saḥ / bṛhataḥ / divaḥ* (RV 6.2.4)[12]
    - *purupriyā ṇa ūtaye* → *puru'priyā / naḥ / ūtaye* (RV 8.5.4)

[11] The *saṃhitā* version and its corresponding *pada-pāṭha* for each of these *mantras* had been taken from the e-source, Vedic Scriptures.

[12] The instances from the Vedic literature are presented throughout the document without their accents. Accents are not considered for both the alignment as well as in the morphological analysis engine and hence ignored but a short account on the importance of accents is provided in Section 4.5.

4. restoration of sounds elided in *saṃhitā-pāṭha*. For example, *yam ī garbham* in the *saṃhitā* is *yam / īm / garbham* in the *pada-pāṭha* (RV 9.102.6)
5. Employing an *avagraha* either for separating various components of a word, like stem and suffixes (*haribhyām* to *hari'bhyām* (RV 1.35.3)), or for separating compound components (*puruvasuḥ* to *puru'vasuḥ* (RV 2.1.5)), or separating a word and *iva* which immediately follows the word (*pragardhinī iva* to *pragardhinī'iva*) (10.142.4)
6. Marking the hiatus with an appended *iti* → *patī* to *patī iti* (1.23.3)
7. Compound words which end in an unchangeable vowel is repeated after *iti* in the *pada-pāṭha* → *vājinīvasū iti vājinī'vasū* (5.74.6)
8. Shortening the vowels lengthened by *pluti* → *acchā vada* to *accha / vada* (5.83.1)
9. Removing the nasal sound used for euphony *śāśadāṁ* to *śāśadā* (1.120.10)
10. Changing the order of words wherever necessary → *śunaścicchepam* to *śunaśchepam / cit* (5.2.7)

In order to obtain the original words, we have to remove the indicators from the *padapāṭḥa*. The *avagraha* is used either as a compound marker or an affix marker. Only one *avagraha* is marked in a particular entry of a *pada-pāṭha*. And there is an order of precedence as to which type of an *avagraha* is employed i.e., where should the *avagraha* be employed: prefix, suffix, compound or when the subsequent word is *iva*. The precedence is: *iva* > compound > suffix > prefix. The *itikaraṇa* is a phenomenon where the word *iti* is inserted either to mark a hiatus or to mark a word that ends in a *pragṛhya*. Sometimes both the *avagraha* and the *iti* can be observed in the *pada-pāṭha* for the same word. The Vedic-specific transformations, elided sounds, the nasal sound and the change in order are observed in the *saṃhitā* while the *pada-pāṭha* has the original versions. The *ṛkprātiśākhya* contains various rules pertaining to these features of the *pada-pāṭha* and also provides exceptions in each of the cases. Additionally, the accents are crucial for disambiguation in various stages, but the current setup of tools do not process the accents and computationally processing accents is a field yet to be explored.

These differences between the *saṃhitā* and the *pada-pāṭha* will play a major role when processing Vedic texts. The traditional sequence of analysis, starting from segmentation, morphological analysis, parsing (sentential analysis) and so on, require in the first place the segmentation. *Padapāṭha* is possibly the first attempt to segmentation in Sanskrit literature. Thus, we intend to use the *pada-pāṭha*, filter it to obtain the *padas* (segmented words) and then generate the unsegmented-segmented pair of *saṃhitā* and *pada*, which can further be used for the subsequent tasks of analysis.

3.2 Vedic Scriptures

The Vedic Scriptures repository contains the following for each of the Vedas:

- *mantra* indices
- *saṃhitā-pāṭha* (with and without accent markers)
- *pada-pāṭha* (with and without accent markers)
- *devatā*, *ṛṣi*, *chandas*
- *svara*,[13]
- commentaries from *Sāyanācārya, Maharshi Dayanand Sarasvati (MDS), Aryamuni, Brahmamuni* and *Shivashankarasharma*

These commentaries contain information regarding the *mantra* (*mantraviṣayaḥ*), word-meanings and interpretation (*bhāvārthaḥ*). MDS' commentary alone has the prose order (*anvaya*) of the *mantras*. We used the unaccented *mantras* from the *saṃhitā* and the unaccented *padas* from the *pada-pāṭha* for our analysis.

[13]This *svara* is a musical note and is different from the accent markers.

3.3 DCS Annotations

The Digital Corpus of Sanskrit (DCS) hosts the *Ṛgveda-saṃhitā* with lexical, morphological and dependency annotations. It also has word senses for some of the mantras. It is available in the CoNLL-U format.[14] The annotations are done for sentences extracted from the original *saṃhitā*. So, there are instances where a mantra having two or more sentences are annotated separately. For example, the second *mantra* from *Ṛgveda-saṃhitā's* first *maṇḍala's* first *sūkta* (1.1.2), is annotated as two separate sentences:

> *agniḥ pūrvebhiḥ ṛṣibhiḥ īḍyaḥ nūtanaiḥ uta*
> *sa devāṁ ā iha vakṣati*

There are also instances where multiple mantras are annotated together in a single sentence. For example, 1.1.7 and 1.1.8 are annotated together:

> *upa tvā agne dive dive doṣāvastar dhiyā vayam namaḥ bharantaḥ ā imasi rājantam adhvarāṇām gopām ṛtasya dīdivim vardhamānam sve dame*

Thus, the definition of a sentence in DCS cannot be uniformly mapped to the mantras. For every sentence, the following annotations are present:

1. word
2. stem / root
3. part of speech category
4. morphological analysis
5. dependency relation
6. link to the lexicon
7. word sense information

3.4 Vedaweb Annotations

Vedaweb hosts the *Ṛgveda* mantras along with their indices (that include the *maṇḍala, sūkta, mantra, pāda* and *pada* or term indices), word, lemma and morphological analysis. This data is available in the TEI format. One major advantage of the Vedaweb version is the indices with the *pāda* marks, which makes the alignment process with the *pada-pāṭha* smoother. Another advantage is the usage of the words according to the *pada-pāṭha* and not according to the *saṃhitā-pāṭha*. This resolves almost all the word-level issues observed in DCS. Thus the 10,552 mantras of the *Ṛgveda* produce a total of 164,767 padas. Of these, 26,573 do not have any morphological analysis marked. These are predominantly indeclinables. The stems are annotated similar to DCS i.e, either one of the base or the derived lemma is used. The morphological analysis is marked based on the following parameters:

- Number: SG, DU and PL
- Case: NOM, ACC, INS, DAT, ABL, GEN, LOC, VOC
- Gender: M, F, N
- Person: 1, 2, 3
- Voice: ACT, PASS, MED
- Tense: PRS, PRF, PLUPRF, FUT, IMPFT, AOR, COND
- Mood: OPT, INJ, SBJV, IND, IMP, PREC
- Participles: PPP, CVB
- Secondary Derivatives: DES

These morphological analysis features can be directly mapped to DCS, except for the ACT, MED voices and DES. In the first case, DCS does not distinguish between active and middle voices. And DCS does not mark the secondary conjugation of a verb like causative, desiderative while Vedaweb marks the verbs with the desiderative suffix.

[14] https://universaldependencies.org/format.html

3.5 Observations on the Alignment

We thus have three versions of the *Ṛgveda-saṃhitā:* Vedic Scriptures (VS), DCS and Vedaweb (VW). An alignment was attempted to prepare a dataset that encompasses the details from the three resources.

3.5.1 Aligning VS and DCS

The sentence-level issues in DCS were discussed earlier. Thus a direct sentence to *mantra* alignment requires additional efforts to merge the sentences wherever the *mantra* has been divided in the DCS, and manual intervention is required when multiple mantras are presented in a single sentence of DCS. So, we relied on the alignment of the *pada-pāṭha* with the segments rather than an alignment of the mantra with the DCS sentences. There were many challenges while aligning the VS *pada-pāṭha* with the DCS segments. Some of these are discussed ahead.

1. DCS does not resolve the terminal sandhis of some words. For example, *punar* (DCS) for *punaḥ* (VS), or *sa* (DCS) for *saḥ* (VS).
2. DCS annotates the *saṃhitā* form of the word predominantly. This means that, words like *sacasva* have their final short vowels changed to their corresponding long *sacasvā* as in the *saṃhitā.* Similar instances can be observed with the dual ending words like *tuvi-jātau* where the *saṃhitā* has *tuvi-jātā.*
3. While a *padapāṭha* presents the components of a compound word in the same entry, DCS provides them in different entries when a compositional meaning is intended. For example, VS has *ratna'dhātamam* while DCS has two entries *ratna* and *dhātamam.* On the other hand, there are instances where a compositional presentation of a compound word in VS is annotated as a non-compositional entry in DCS. For example, *citraśravaḥ'tamaḥ* (VS) vs *citraśravastamaḥ* (DCS).
4. In some cases, the *pada-pāṭha* proposes a non-compositional representation, especially when there are more than two components, while DCS sticks to the compositional representation. For example, *surūpa'kṛtnum* (VS) vs *su-rūpa-kṛtnum* (DCS). This is because of a constraint in the *pada-pāṭha* that an entry should have only one *avagraha*, and the compound boundary marker has a higher preference than a prefix marker.
5. Sometimes, two entries of a *pada-pāṭha* are combined into a single entry in DCS. For example, *parā / ihi* (VS) vs *parehi* (DCS). It is not trivial to map automatically such cases.
6. In some cases where the preverbs are not joined with their corresponding verbs, and these involve in a sandhi with one of their neighbouring words, DCS skips such preverbs. For example, *indra / ā* (VS) vs *indra* (DCS).
7. In some cases, the pada-pāṭha entry is incorrect. These had to be manually checked with the help of a valid source text.[15]

These word-level differences hinder the alignment of the DCS annotations with the VS. The stem and morphological analysis of the DCS have limitations which have to be addressed using an alignment with other morphological analysis tools like SH and SCL. Table 1 shows the summary of the alignment between the *pada-pāṭha* of VS and the entries of DCS. We observe that the unmatched 8,348 entries of the *pada-pāṭha* are distributed across the 4,911 mantras. Possibly these correspond to the 9,403 unmatched entries of DCS.

3.5.2 Aligning VS and Vedaweb

An alignment was attempted between the *pada-pāṭha entries* of VS and VW where comparisons were done based on: (1) pada-pāṭha indices, (2) word (stripping the accents as VW data did not contain the accents), and (3) similarity between the VW and VS based on approximate Levenshtein edit distance.[16] The observations on the alignment of the VS *pada-pāṭha* and VW

[15] We used the *Ṛgveda-saṃhitā* with the commentary of *Śāyanācārya*, published by Vaidik Samshodhana Mandal.

[16] Such similarities introduce errors based on characters and ignore certain minute differences. But we kept this as an approximate measure to understand the differences.

	Mantra	Padapāṭha
Number of entries in VS	10,552	163,396
Number of entries extracted from DCS	10,527	169,955
Missed in DCS	25	337
Matched	10,527[1]	154,496[2]
Unmatched in VS	4,911	8,348
Unmatched in DCS	5,648	9,403

[1] This denotes partially matched mantras. Only 5,616 mantras have a complete match.

[2] This includes 145,523 entries mapped directly and 8,973 mapped after merging the components of a compound present as multiple entries in DCS.

Table 1: Alignment of Vedic Scriptures Padapāṭha with DCS entries

entries are as follows:

1. The number of entries differ in the two systems: 163,396 (VS) vs 164,766 (VW). It was observed that some of the *pada-pāṭha* entries contain multiple words. For example, in the *mantra* 10.184.3, the last *pada-pāṭha* entry corresponds to *havāmahe daśame māsi sūtave*, where each of them have to be considered as separate *pada-pāṭha* entries. Since there were many such entries which required further validation, we relied on the words rather than the *pada-pāṭha* entries of VW.
2. The character *ḷ*, which can be used interchangeably with *ḍ*, introduced ambiguities and was thus converted to *ḍ* across both the datasets for use in the further stages of the alignment, since SH and SCL process only *ḍ* and not *ḷ*.
3. Terminal sandhi needs to be resolved in some cases: *viśvatas* (VW) vs *viśvataḥ* (VS).
4. Some cosmetic corrections had to be done. For example *gachati* to *gacchati*, *acha* to *accha*, etc.
5. VS has the *iva* attached to the word and VW has a separate entry for *iva*. There are 1,024 occurrences of *iva* in VW and 1,021 occurrences in VS. The difference in the number of *pada-pāṭha* entries between VS and VW could be attributed to these additional entries of *iva* as well. Thus, for the alignment, these *iva* were attached to their previous term in VW to match with the VS.
6. Similar to DCS, the preverb (*ā*) has not been considered for some terms. For example, *ā / omāsaḥ* (VS) vs *omāsaḥ* (VW)
7. In the case of compounds, where VS has an *avagraha*, SCL's sandhi module is incorporated to perform the sandhi between the components. In some instances, SCL does not handle the Vedic-specific sandhi constructions. For example, *su-stutim* in the *pada-pāṭha* becomes *suṣṭutim* in the *saṃhitā* and VW provides the *saṃhitā* form. In such cases, either the sandhi engine should be augmented to handle Vedic sandhi, or these instances have to be manually mapped.
8. Table 2 provides the results of the alignment. The 1,032 unmatched entries have to be manually analysed by verifying an authentic source.[17]

3.5.3 Aligning DCS and Vedaweb

As the alignments between VS-DCS, and VS-VW have already been done, their results were aligned to produce a unified VS, VW, DCS dataset. The results are presented in table 3. The aligned dataset contains annotations for 154,269 *pada-pāṭha* entries.

[17] This manual verification is in process and the final database will be updated when the verification gets completed.

	Mantra	Padapāṭha
Number of entries in VS	10,552	163,396
Number of entries extracted from VW	10,550	163,742
Matched	10,550[1]	162,382
Unmatched in VS	900	1014
Unmatched in VW	1,025	1,360

[1] 9,614 mantras have a complete match and the remaining are partially matched mantras.

Table 2: Alignment of Vedic Scriptures Padapāṭha with Vedaweb entries

	Mantra	Padapāṭha
Number of entries from VS-DCS	10,527	154,496
Number of entries from VS-VW	10,550	162,364
Matched	10,525[1]	154,269
Unmatched VW in DCS[2]	4,679	8,095
Unmatched DCS in VW[3]	157	227

[1] This shows partially matched *mantras*.

[2] This indicates those VW entries which couldn't be found in DCS.

[3] This indicates those DCS entries which couldn't be found in VW.

Table 3: Alignment of DCS with Vedaweb entries using the results of the previous two alignments

3.5.4 SH and SCL annotations

SH and SCL morphological analyzers are lexicon-driven and paradigm-based analyzers which use finite-state automata for analysis. The VS *pada-pāṭha* was transformed into the padas which were run on these two systems. Converting a *pada-pāṭha* into its corresponding pada involved various measures that remove or transform the special indicators like *itikaraṇa* or *avagraha*. Table 4 shows the performance comparison of the two morph analyzers over the *Ṛgveda* words.

	SH	SCL
Number of entries	163,396	163,396
Number of unique entries (with accents)	33,941	33,941
Number of unique entries (without accents)	30,633	30,633
Morph analysis obtained	18,639	14,205
Unrecognized	11,994	16,428

Table 4: Performance of the Morphological Analyzers in the *Ṛgveda Padapāṭha*

The morphological annotations of SH and SCL follow a different approach. They provide the base and derived stems when both are available. And also provide various other morphological features like voice, class, secondary conjugation, etc. But a major disadvantage is their inability to handle the peculiar features of Vedic Sanskrit. Since both the systems (SH and SCL) are lexicon-dependent, updating their lexicon will definitely reduce the number of unrecognized words. But, we will still have a significant number of words for which SH or SCL fail to produce their morphological analyses. And both provide all possible morphological analyses for a given input, and contextual morphological analysis can only be obtained in the subsequent stages of processing.

The DCS morphological analysis obtained earlier (for 154,269 entries) were aligned with the corresponding possible SH and SCL analyses and the observations are recorded in table 5. We observe that for 72.6% of the aligned *pada-pāṭha*,[18] the DCS analysis was aligned with a single analysis of SH, and for 74.8% of the aligned *pada-pāṭha*,[19] a single analysis of SCL was aligned with the DCS analysis. The overall aligned dataset contains the details from VS, VW, DCS, SH and SCL.

	SH	SCL
Alignment with DCS	112,029	115,497
No analysis	42,225	38,757
Mismatches with DCS	15	15

Table 5: Alignment of the DCS Morphological Analyses with the analysis from SH and SCL of the Ṛgveda Padapāṭha

4 Challenges in processing the *pada-pāṭha*

One of the limitations of SH's analyzer is its inability to recognize words with certain secondary suffixes (*taddhita* forms). And not all the primary derivations are recognizable. Also, it has a limited lexicon which is in continous development and certain words which are not in the lexicon go unrecognized. The following is an account on the challenges observed in the unrecognized words, at various levels of word generation.

While some of the challenges exist due to the differences between Classical and Vedic Sanskrit, some are due to the changes in the *pada-pāṭha*, and some challenges are due to the limitations of the tools used. SH uses vocabulary from various literature sources and word-generation rules from *Pāṇini's Aṣṭādhyāyī* for its morphological analysis. It mainly considers the constructions and vocabulary from classical Sanskrit literature while certain forms and stems are available only in the Vedic context. For example, the subjunctive mood (*leṭ-lakāra*) is not handled by SH, while such words are found only in Vedic scriptures. While there are multiple infinitives in Vedic, we find only one in Classical (with the suffix *tumun*).[20] For some of the words not recognized by these tools, the dictionaries of SH (Monier-Williams and Sanskrit-French) and those of Samsādhanī (Sanskrit-Hindi Apte, MW, Sanskrit-German) have to be updated. For some words, the paradigms have to be updated to incorporate the Vedic constructions. Thus we describe ahead our observations on some of the challenges and the methods we deployed to handle them.

4.1 upasarga

In Vedic Sanskrit, a preverb (*upasarga*) and its corresponding verb are encountered separately, but in Classical Sanskrit, they are always together. The *Nirukta* of *Yaska* gives a list of all preverbs occuring in the Vedic literature. *Nirukta* and its ancillary text *Nighaṇṭu* are the primary sources of evidence for etymological analysis. *Nighaṇṭu* enlists the words that occur in the Vedic literature.[21] And *Nirukta* presents the rules to disambiguate them.

Nirukta enlists 20 upasargas and mentions that they are used to indicate different kinds of special meanings from the same root.[22] Nirukta also states the view of *Śākatāyana* that *upasargas* are indicative (*dyotaka*) rather than denotative (*vācaka*) and also that they cannot present a clear meaning when detached from verbs or nouns, but only express a subordinate sense of nouns

[18] 68.5% of the overall *pada-pāṭha* entries

[19] 70.7% of the overall *pada-pāṭha* entries

[20] *tumarthe se-sen-ase-asen-kse-kasen-adhyai-adhyain-kadhyai-kadhyain-śadhyai-śadhyain-tavai-taveṅ-tavenaḥ - Aṣṭādhyāyī 3.4.9* gives a list of suffixes used in the Vedic literature in the sense of *tumun*.

[21] *Nighaṇṭu* is not an exhaustive list of all the words present in the entirety of Vedic literature but contains a huge list of words whose etymological and morphological analyses are ambiguous.

[22] *nānāvidha-viśeṣa-artha-pradhāna*

and verbs.[23] And also that according to Gārgya, *upasargas* have various meanings (even when they are detached from a noun or verb), each of which implies a modification in the meaning of the corresponding Noun and Verb (Sarup, 1967).[24]

For our analysis, though, it is necessary to analyse preverbs even if they are independently existing in the *pada-pāṭha.* Joining the preverbs with their corresponding verb or noun requires additional information like relationship between the words, which can be done only in the subsequent stages. Since SH and SCL do not analyse preverbs independently, a new category of preverbs was introduced in the same format as that of SH. DCS does not annotate any morphological analysis for indeclinables and preverbs. The information regarding indeclinables, and preverbs, can be obtained from DCS' dictionary. With the help of these, we introduced a new category ('prev.') for preverbs.

4.2 itikaraṇa

itikaraṇa is one of the phenomenon where the word *iti* is added to the *pada* on special occasions. The *itikaraṇa-lakṣaṇa* gives two kinds of situations where *iti* is added to a *pada*: (1) *pada* is not repeated after *iti* and (2) *pada* is repeated after *iti.* For example, *akṣī iti* and *gopatī iti go'patī*, respectively. There are seven scenarios where such a phenomenon occurs:

1. after a word with final *pragṛhya* vowel[25] (*o*; dual endings *ī, ū, e*, locative endings in *ī* or *ū, amī, asme, yuṣme, tve*). *Pragṛhya* vowels remain unchanged if placed before a vowel (indicating an absence of vowel sandhi). Vocatives with final *o* are *pragṛhya* in the *pada-pāṭha* only. Examples: *agnī iti*; *śatakrato iti śata-krato*, *tanū iti*, *asme iti.*
2. words ending in *aḥ* or *āḥ* in which the final *visarjanīya* comes from *r* or *s.* This insertion of *iti* is done only when the *visarjanīya* is placed before *r*, an unvoiced consonant or a pause. Examples: *punariti, kariti kaḥ, svariti svaḥ.*
3. particle *īm* when the final *m* is dropped in the *saṃhitā (īmiti).*
4. the particle *u* which is *pragṛhya* in the *pada-pāṭha* only (*ūṁiti*).
5. ten verbs ending in *uḥ, eḥ* and *oḥ.* Examples: *ūvurityūvuḥ, pīperiti pīpeḥ, tūtoriti tūtoḥ.*
6. three nouns ending in a *visarjanīya* which comes from *s*: *rathyebhiriti rathyebhiḥ; praceta iti pra-cetaḥ, sta iti staḥ.*
7. Seven words which do not end in a *visarjanīya*: *gdheti gdha, ta iti te, namasyanniti namasyan, pranapād iti pra-napāt, vargiti vark, syasveti syasva, hanniti han.*

The *itikaraṇa* text, from the *Vedalakṣaṇa-Granthas*,[26] enlists all the *padas* in each of the cases and these were compared with the VS *pada-pāṭha.* There are 682 *padas* with *iti* in *Ṛgveda* and each of these were stripped of the *iti* and the additional word, and then checked for morphological analysis from SH and SCL. There were 55 cases where the words from the *itikaraṇa* did not match any of the words in the *pada-pāṭha.* There were 165 cases where the morphological analyses was not obtained automatically from SH or SCL. For these two cases, manual validation and manual annotation of the morphological analysis was done. For the remaining 462 words, both the validation and morphological analysis resulted in a success. SH and SCL produce all possible morphological analysis of a given word. To arrive at the intended morphological analysis according to the context, one has to manually pick the required analyses. On the other hand, with the annotations from DCS and VW, for some of the *pada-pāṭha* entries, the SH and SCL morphological annotations were aligned with the DCS annotations to select the most appropriate analysis.

[23] *na nirbaddhā upasargā arthānnirāhuriti śākaṭāyanaḥ. nāmākhyātayostu karmopasṃyogadyotakā bhavanti.*

[24] *uccāvacāḥ padārthā bhavantīti gārgyaḥ. tad ya eṣu padārthaḥ prāhurime taṃ nāmākhyātayorarthavikaraṇam.*

[25] *Pragṛhya* is a vowel not liable to the sandhi rules.

[26] This was obtained from the analysis on various *Vedalakṣaṇa-granthas* available at: `https://sites.google.com/view/vedalakshana`

4.3 avagraha

Generally, an *avagraha* is used to indicate two types of *sandhi*: (1) when a word starting with the vowel *a* follows a word that ends in a *visarga*, and (2) when a word starting in a vowel follows a word ending in either the same vowel or a *savarṇa* of the vowel. In the *pada-pāṭha*, which has only the segmented words and there is no possibility of any *sandhi*, an *avagraha* has a special meaning to denote certain information regarding the *pada*. There are four cases where an *avagraha* is inserted:

1. separating the stem from suffixes
2. splitting the compound components
3. separating a word from *iva* which immediately follows the word
4. separating a prefix from the verb or noun

For the first case, one way to handle is to detect all possible suffixes, merge them with their stem-forms and then generate their morphological analysis. For the second case, each of the components of the compound has to be checked for its morphological analysis. The third case can be handled by separating the *iva* from the *pada*.

We replaced the *avagraha* with a "-" to make sure that the tools do not misinterpret it as the *avagraha* because of a *sandhi*. And during the comparison, we maintain two forms viz. sandhied and hyphenated,[27] and then match both of these with the words of DCS or VW. Further we extract the morphological analysis of both of these forms.

4.4 Special cases like nasal sound

In the *mantras*, the pronunciation of the character *m* has many varieties. The primary difference in the writing is reflected in the change of *m* to *ṃ* (*anusvāra*), even in Classical Sanskrit. The pronunciation brings forth another variety with the influence of the adjacent characters. The pronunciation when the *anusvāra* is followed by a *v*, differs from when followed by *y*, or any character of the *p-varga*. In addition to this, in some *mantras*, a nasal sound (ṁ) is used in place of *m* / *ṃ*. During recitation, it is pronounced as *gum*, and sometimes it is accented too, increasing the number of possible sounds from a single character *m*. This nasal sound also appears at the end of a word where the final character of the word belongs to one the five nasal characters (*ṅ, ñ, ṇ, n, m*). Also, when it is followed by one of the *s* / *ś* / *ṣ*, the pronunciation changes from *gum* to *gus* / *guś* / *guṣ*. Thus its usage and variety are prevalent in the *saṃhitā-pāṭha* mainly due to the occurrence of *sandhi*. And in the *pada-pāṭha*, only the intra-word modifications of the nasal sound remain. For our analysis, we replace the nasal sound marker with one of the five nasal characters, where the possibility of *m* is higher than the rest.

4.5 Accent (*svara*):

Accents provide both grammatical and semantic information of the words but with the loss of accent in Classical Sanskrit, ambiguity has increased. The nature of the ambiguity in non-accented forms can be observed from the statistics extracted from our base data. In the *Ṛgveda* alone, the number of unique terms ignoring the accents is 30,633 while the number of unique terms with their accents is 33,941. This difference of 3,308 is distributed across 2,993 terms. So 27,640 terms have a single accented form, while 2,993 have multiple accented forms. The maximum number of accented forms for a word reaches upto 6 for three words (*marutaḥ, indra* and *agne*). Eight words have five accented forms. 31 have 4 accented forms. 218 words have three accented forms and the remaining 2,734 have two accented forms. Generally, the number of accented forms of a word is proportional to the length of the word. But, in this case, we also find smaller words with accents at different positions leading to multiple accented forms.

[27] We used Saṃsādhanī's sandhi engine to perform sandhi between the components. We understand that there are differences between Classical Sanskrit and Vedic Sanskrit on how sandhi occurs. But as there is yet to be an engine that handles Vedic sandhi rules, we relied on the existing sandhi engine that would address most of the cases.

There are different kinds of accent markers. For example, *Ṛgveda* has three *svaras*: *udātta, anudātta, svarita. Yajurveda* has four: *udātta, anudātta, svarita, dīrgha-svarita.* The pronunciation of *svarita* of *Yajurveda* differs from that of *Ṛgveda.* And due to the existence of various branches, we find differences between what is referred to as *udātta* in one Veda, is referred to as *svarita* in another Veda. On the other hand, the *Sāmaveda* initially consisted of 3 *svaras* similar to *Ṛgveda*, but later it expanded the three basic *svaras* into seven. These are represented as numbers atop the characters. While 70% of *atharvaveda-saṃhitā* has lost its *svaras*, we can find a unique representation of *svaras* in the remaining 30%. In addition to the *udātta, anudātta* and *svarita*, there is a *jātya-svarita* which is recited like a *svarita* depending on whether a short or long vowel generated by the *sandhi.*

Conversion of a *saṃhitā-pāṭha* to its corresponding *pada-pāṭha* involves resolving *sandhi* with respect to words as well the *svaras* of the words. As there are rules of *sandhi* for non-accented words, there are several rules of *sandhi* for accented words too. Most of the *sandhi* (external) are categorised into the three: *praśliṣṭa, abhinihita* and *kṣaipra.* VS has both accented and non-accented entries for *saṃhitā-pāṭha* and *pada-pāṭha.* Since our analysis depends on SH, SCL and the existing data in DCS and Vedaweb, all of which do not consider accents in their analysis, we have ignored the *accents* in the current setup.

5 Vedic Morphological analysis Engine

Taking insights from the alignments of VS, VW and DCS, and the morphological analyses produced by the tools SH and SCL, we have come up with a morphological analysis engine that generates the possible morphological analyses of a *pada-pāṭha.* It extracts the morphological analyses from SH, SCL and the pre-existing analyses annotated by platforms like DCS and VW. Here we describe the architecture of this engine.

1. Preprocessing: The input is required to be in one of the notations: Devanagari, IAST, WX and SLP. The first step involves pre-processing the input where the accent markers are removed as the SH and SCL tools do not process them.[28] E-versions of most texts may contain non-unicode characters and these are also removed. There are some special characters used in the Vedic texts, for example *ḷ* which is not processed by the SH and SCL engines, hence converted to its alternate *ḍ*. Another special character is the nasal sound (*gum*), which is converted to *m*. Finally the avagraha is replaced with a hyphen to avoid being considered as the avagraha because of *pūrvarūpa* or *savarṇadīrgha sandhiḥ.*

This preprocessing step also involves a sub-module that handles the *itikaraṇa.* The list of all the occurrences of the *iti* extracted earlier was used here. Patterns were extracted from these to split the iti from the pada. The two types of iti: with and without repetition are handled here. The terms with avagraha are further sandhied using SCL's sandhi engine. This sandhi joiner does not produce all Vedic sandhi occurrences. A selected few cases which were repeatedly occurring were proposed as exceptions and the rest were run on the sandhi engine. Thus, this preprocessing module produces three outputs: segmented (where the iti is split), sandhied (to remove the avagraha) and hyphenated (the avagraha is retained here as a hyphen).

2. SH morphological Analysis: The SH segmenter can also be used as a morphological analyser where given a word, it produces all possible morphological analyses along with marking the compound boundaries. SH returns a JSON object with the segmentation and all possible morphological analyses as the features. This output is processed further to produce the results in a standard format. For example:


```
{
    "input": "hitam",
    "status": "success",
    "segmentation": ["hitam"],
```


[28] But the accented *pada-pāṭha* entries are not removed entirely as the various functionalities of accents are useful in subsequent stages of processing like compound analysis, sense disambiguation and parsing.

```
    "morph": [
        {
            "word": "hitam",
            "stem": "hita#1",
            "root": "hi#2",
            "phase": "Kric",
            "derivational_morph": "pp.",
            "inflectional_morphs": [
                "n. sg. acc.", "n. sg. nom.", "m. sg. acc."
            ]
        },
        {
            "word": "hitam",
            "stem": "hita#2",
            "root": "dh\={a}#1",
            "phase": "Kric",
            "derivational_morph": "pp.",
            "inflectional_morphs": [
                "n. sg. acc.", "n. sg. nom.", "m. sg. acc."
            ]
        }
    ],
    "source": "SH"
}
```


Each of the morphological analyses contains the word, its stem with or without the root, phase (or part of speech), derivational and inflectional morphological analysis. The notations are as produced by the SH engine.

3. SCL morphological analysis: The third step is where Saṃsādhanī's morphological analyser is used to produce the possible morphological analyses. SCL uses Apertium's *lttoolbox* package for its morphological analysis. The results are thus produced in an XML-like pattern which are converted to the JSON format as described earlier. A mapping is established between the SCL and SH representations to convert the morphological analysis. An advantage in this conversion is that, majority of the tags proposed in SCL are available in SH too. SCL additionally produces analysis of various *kṛt* suffixes like *ghañ*, *tṛc*, *lyuṭ*, etc. and a few *taddhita* suffixes like *matup*, *vat*, *tva*, etc. which are not produced by SH.

4. DCS analysis: All the morphological annotations proposed by DCS from the *Ṛgveda* and *Atharvaveda* were collected and converted to the SH format as prescribed above. Since SH produces more information like conjugation, class, etc, this conversion could result into multiple SH possibilities for a single DCS morphological analysis. But such additional features are kept hidden to avoid unnecessary duplicates. In case both SH and SCL fail in producing the results, this list of DCS words and their morphological analyses helps in assigning the possible analysis.

5. Morph-merger: The final step involves comparisons of the analysis from SH, SCL and DCS.[29] The analysis is obtained for the sandhied and the hyphenated versions of the *pada-pāṭha* from each of the three systems. In total, we have six morphological analysis results from which the SH analysis on the sandhied input is given higher preference followed by its hyphenated version. This is followed by the SCL analysis on the sandhied input and then the hyphenated input. Finally, DCS is considered similarly. An alignment of the possible analyses from SH and SCL with the DCS analysis was done to produce a single morphological analysis that merges the results of the three systems.

[29] Vedaweb analysis was not considered as it is predominantly similar to DCS analysis as observed in a sample set of 10 mantras.

Dataset details: With the help of the alignments between VS, DCS and VW, we were able to align a majority of the *Ṛgveda-pada-pāṭha* entries across the three annotations. Along with the analyses generated from the above engine, the alignment gave us a dataset comprising of the following:

1. Mantra Index (according to both the *Maṇḍala* order and the *Aṣṭaka* order)
2. Mantra (with and without Svara)
3. *Pada-pāṭha* (with and without Svara)
4. Other details from Vedic Scripture like *chandas, ṛṣi, devatā* and translations
5. For each of the padas of the *Padapāṭha*:
   (a) Pada index
   (b) Pāda annotation
   (c) Vedaweb stem
   (d) Vedaweb morphological analysis
   (e) DCS stem
   (f) DCS morphological analysis
   (g) SH morphological analysis (if available)
   (h) SCL morphological analysis (if available)

The aligned dataset consists of the annotations for 154,269 entries of the *pada-pāṭha* where 5,408 *mantras* have been completely aligned and 5,117 *mantras* have been partially aligned. Aligning the remaining *pada-pāṭha* (approx. 9,127) requires further processing.[30]

6 Conclusion

The present work discusses various challenges faced in an attempt to process Vedic Sanskrit computationally. The primary motivation was to create a database for the *Ṛgveda-saṃhitā* which has for each of the mantras, its *pada-pāṭha*, *padas* (segmented words), *pāda* information (metrical unit), relevant information regarding the mantras like the *devatā, chandas, ṛṣi,* translation, etc. and most importantly the lexical and morphological analysis of each of the padas. With the availability of the *saṃhitā* and the *pada-pāṭha* in the Vedic Scriptures platform and the lexical and morphological information from the Vedaweb and the Digital Corpus of Sanskrit, an alignment was carried to map the details across the three platforms to produce a single source of information. The morphological analysis from SH and SCL platforms were also extracted for each of the padas. In this process, the challenges present in the *pada-pāṭhas* were discussed and the possible solutions to handle them were also provided.

In the current scenario, the alignment between VS and DCS was done for 94.5% of the *pada-pāṭha* entries, where 46.5% of the *mantras* were completely aligned and the remaining were partially aligned. The unaligned *mantra* to unaligned *pada-pāṭha* ratio (4,911 / 8,348) shows that on an average, every *mantra* has almost two unaligned words. The alignment between VS and VW showed promising results, where close to 99.3% of the *pada-pāṭhas* were aligned and more than 80% of the *mantras* were completely aligned. The unaligned mantra to unaligned pada-pāṭha ratio (900 / 1014) shows that at least one word per *mantra* went unaligned in these *mantras*. However, to build an error-free system or a resource, it requires multiple validations, involving both human as well as computational efforts. There are collections of texts called *Vedalakṣaṇa-granthas* which provide an exhaustive analysis about the phonological and morphological features of the *saṃhitā, pada-pāṭha, kramapāṭha,* etc. which are valuable resources

[30]The alignment results and the aligned dataset are available here: `https://github.com/SriramKrishnan8/svarupa_alignment.git`.

for validating and preserving the Vedic texts. These *lakṣaṇa-granthas* can be used further on the VS, VW and DCS for validating their annotations.

The two morphological analysers: SH and SCL, produced morphological analysis for 68.5% and 70.7% of the overall *pada-pāṭha* entries, respectively. These tools are in continuous development and the performance is expected to increase when their lexicons and paradigms are updated to include Vedic forms. However, the morphological analysis in context can only be produced in the subsequent tasks of parsing or sentential analysis.

The Vedic morphological analysis engine is developed with a view to create a framework that processes Vedic texts. This engine handles the pecularities of the *pada-pāṭha* and can be used to extract the morphological analysis from SH, SCL and the annotations of DCS.[31] The DCS annotations are helpful in providing the analysis for similar Vedic forms in other Vedas, especially *Sāmaveda*. The overall engine and the individual modules are available publicly as follows:

- Vedic Morphological Analysis Engine:
  `https://github.com/SriramKrishnan8/svarupa_morph_analysis.git`
- SH Morphological Analyser:
  `https://github.com/SriramKrishnan8/vedic_morph_analyser_sh.git`
- SCL Morphological Analyser:
  `https://github.com/SriramKrishnan8/scl_morph_interface.git`
- DCS Morphological Analysis:
  `https://github.com/SriramKrishnan8/dcs_morph_analysis.git`

7 Acknowledgement

We would like to thank Dr. Bhagyalatha Pataskar and Dr. Jayashree Sathe for providing us valuable inputs regarding the grammar of Vedic Sanskrit, the structure of Vedas, the details in the *pada-pāṭha*, and many other such information that helped us direct towards our goal. The present work is a part of the Project titled 'Svarupa' and we would also like to thank Shri Arimilli Ramana Rao for conceptualizing the project.

References

Erica Biagetti, Oliver Hellwig, and Sven Sellmer. 2023. Hedging in diachrony: the case of Vedic Sanskrit iva. In Daniel Dakota, Kilian Evang, Sandra Kübler, and Lori Levin, editors, *Proceedings of the 21st International Workshop on Treebanks and Linguistic Theories (TLT, GURT/SyntaxFest 2023)*, pages 21–31, Washington, D.C., March. Association for Computational Linguistics.

Braj Bihari Chaubey. 1972. *Vaidika Svara Bodha*. Vaidik Sāhitya Sadan, Hoshiyarpur, Punjab.

Pawan Goyal and Gérard Huet. 2016. Design and analysis of a lean interface for Sanskrit corpus annotation. *Journal of Language Modelling*, 4(2):145–182, Oct.

Pawan Goyal, Gérard Huet, Amba Kulkarni, Peter Scharf, and Ralph Bunker. 2012. A Distributed Platform for Sanskrit Processing. In *Proceedings of COLING 2012*, pages 1011–1028, Mumbai, India, December. The COLING 2012 Organizing Committee.

Oliver Hellwig, Heinrich Hettrich, Ashutosh Modi, and Manfred Pinkal. 2018. Multi-layer annotation of the rigveda. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

[31] The engine is being used to extend this work towards developing datasets for the other Vedas.

Oliver Hellwig, Salvatore Scarlata, Elia Ackermann, and Paul Widmer. 2020. The treebank of vedic Sanskrit. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5137–5146, Marseille, France, May. European Language Resources Association.

Oliver Hellwig, Sven Sellmer, and Kyoko Amano. 2023. The Vedic corpus as a graph. an updated version of bloomfields Vedic concordance. In Amba Kulkarni and Oliver Hellwig, editors, *Proceedings of the Computational Sanskrit & Digital Humanities: Selected papers presented at the 18th World Sanskrit Conference*, pages 188–200, Canberra, Australia (Online mode), January. Association for Computational Linguistics.

Oliver Hellwig. 2009. SanskritTagger, a stochastic lexical and pos tagger for Sanskrit. *Lecture Notes in Artificial Intelligence*, page 266–277, Jan.

Oliver Hellwig, 2010–2021. *The Digital Corpus of Sanskrit (DCS)*.

Gérard Huet and Amba Kulkarni. 2014. Sanskrit linguistics web services. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 48–51.

Gérard Huet. 2003. Lexicon-directed Segmentation and Tagging of Sanskrit. In *XIIth World Sanskrit Conference, Helsinki, Finland. Final version in Themes and Tasks in Old and Middle Indo-Aryan Linguistics, Eds. Bertil Tikkanen and Heinrich Hettrich.*, pages 307–325, Delhi, August. Motilal Banarsidass.

Gérard Huet. 2005a. A functional toolkit for morphological and phonological processing, application to a sanskrit tagger. *J. Funct. Program.*, 15(4):573–614, jul.

Gérard Huet. 2005b. A Functional Toolkit for Morphological and Phonological Processing, Application to a Sanskrit Tagger. *J. Functional Programming*, 15,4:573–614.

Gérard Huet. 2009. Sanskrit Segmentation. In *Proceedings of the South Asian Languages Analysis Roundtable XXVIII*, October.

Sriram Krishnan, Amba Kulkarni, and Gérard Huet. 2023. Validation and Normalization of DCS corpus and Development of Sanskrit Heritage Engine's Segmenter. In *Proceedings of the Computational Sanskrit & Digital Humanities: Selected papers presented at the 18th World Sanskrit Conference*, pages 38–58, Canberra, Australia (Online mode), January. Association for Computational Linguistics.

Sriram Krishnan, Amba Kulkarni, and Gérard Huet. 2024. Normalized dataset for sanskrit word segmentation and morphological parsing. *Language Resources and Evaluation*, Aug.

Amba Kulkarni and Devanand Shukl. 2009. Sanskrit morphological analyser: Some issues. *Indian Linguistics*, 70(1-4):169–177.

P. K. Narayana Pillai. 1941. The Ṛgveda padapĀṬha—a study with special reference to the Ṛgveda prĀtiŚĀkhya. *Bulletin of the Deccan College Research Institute*, 2(3/4):247–257.

Lakshman Sarup, editor. 1967. *The Nighaṇṭu and The Nirukta, The oldest Indian treatise on etymology, philology and semantics*. Motilal Banarsidass, Delhi.

Peter Scharf and Malcolm Hyman. 2009. *Linguistic Issues in Encoding Sanskrit*. Motilal Banarsidass, Delhi.

Sven Sellmer and Oliver Hellwig. 2024. The Vedic compound dataset. In Archna Bhatia, Gosse Bouma, A. Seza Doğruöz, Kilian Evang, Marcos Garcia, Voula Giouli, Lifeng Han, Joakim Nivre, and Alexandre Rademaker, editors, *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 50–55, Torino, Italia, May. ELRA and ICCL.