ISSN: 1672 - 6553

JOURNAL DYNAMICS AND CONTROL

VOLUME 9 ISSUE 6: 82 - 87

COMPUTATIONAL CHALLENGES IN ANCIENT LANGUAGE MODELING: THE CASE OF THE RIGVEDA

[1]Rohit D.S., [2]Vijay Kumar K., [3]Vikas S., [4]Vivek Goutam, [5]Purnima S.M.

[1,2,3,4] Scholars, [5]Assistant Professor,

Department of Computer Science and Engineering, R.N.S. Institute of Technology, Bengaluru, India

JOURNAL OF DYNAMICS AND CONTROL

DOI: https://doi.org/10.71058/jodac.v9i6007

COMPUTATIONAL CHALLENGES IN ANCIENT LANGUAGE MODELING: THE CASE OF THE RIGVEDA

[1]Rohit D.S., [2]Vijay Kumar K., [3]Vikas S., [4]Vivek Goutam, [5]Purnima S.M.

1,2,3,4 Scholars, [5]Assistant Professor,

Department of Computer Science and Engineering, R.N.S. Institute of Technology, Bengaluru, India

1rn22cs132.rohitds@rnsit.ac.in[1], 1rn22cs182.vijaykumark@rnsit.ac.in[2], 1rn22cs183.vikass@rnsit.ac.in[3], 1rn22cs185.vivekgoutam@rnsit.ac.in[4], purnimamittalkod@rnsit.ac.in[5]

***Abstract*—The Rigveda, composed over three millennia ago, is among the earliest known literary texts and an invaluable source of Indo-European linguistic, cultural, and philosophical heritage. With 1,028 hymns written in Vedic Sanskrit which is a very archaic form of Sanskrit yet to be fully deciphered, its structure poses considerable challenges to linguists and is yet to be subject to computational analysis. As Natural Language Processing (NLP) evolves, it becomes an essential tool in interpreting such ancient texts, helping us look at this ancient text in a way like never before. This survey explores the state-of-the-art NLP approaches applied to Vedic literature, focusing on the Rigveda. We review efforts in corpus creation [3], [4], [10], [21], [26], [27], morphological analysis [1], [6], [12], syntactic and semantic parsing [3], [5], [11], [13], [23], and the use of deep learning models [14], [15]. We also examine the specific linguistic challenges in processing Vedic Sanskrit, such as compound disambiguation and poetic syntax [7], [9]. Through critical analysis, we highlight gaps in current research and suggest future directions, including semantic web integration [19], cultural heritage preservation, and advanced question-answering systems [18]. This paper aims to provide a roadmap for computational philologists and NLP researchers venturing into Vedic studies and to revitalize curiosity on studying the RigVeda.**

***Index Terms*—Rigveda, Natural Language Processing, Vedic Sanskrit, Computational Linguistics, Ancient Text Mining.**

I. INTRODUCTION

The Rigveda is widely acknowledged as the oldest extant text in the Indo-European language family, is thought to be composed between 1500 and 1200 BCE. It comprises over a thousand hymns (suktas) organized into ten books (mandalas), featuring a complex poetic meter and a pitch accent system preserved from PIE, making Vedic Sanskrit one of the few languages along with Ancient Greek that best preserves this feature. Its language—Vedic Sanskrit—holds great significance in both Indo-European linguistics and in oriental studies of the Indian subcontinent. As such, the Rigveda represents not only an invaluable philological artifact but also a formidable challenge for computational analysis due to its archaic nature, and the obscurity of the language not to mention the text's semantic richness, and poetic structure very distinct from that of later classical works.

With the advent of modern computational linguistics and Natural Language Processing (NLP), there has been a surge in efforts to analyze ancient languages. NLP offers scalable methods for extracting linguistic patterns, semantic relationships, and structural annotations from large corpora. In the context of the Rigveda, these tools can facilitate deeper philological inquiry, linguistic reconstruction, and reveal patterns in a way that manual analysis just cannot.

However, NLP methods developed for contemporary languages often fail when applied to Vedic Sanskrit due to unique linguistic features such as extensive use of sandhi (euphonic conjunctions) where spoken aberrations are written down unlike in English, rich morphological inflection, and ambiguity in interpretation due to the sheer antiquity of the language [22]. These issues are compounded by the scarcity of annotated corpora and the oral tradition of transmission, which often omits punctuation and standard textual delimiters [1], [2], [9] but more importantly it obscures word roots and orthography.

Despite these challenges, notable progress has been made. Digital corpora such as the Digital Corpus of Sanskrit (DCS), The Metrically Restored Rigveda, the GRETIL project and The Sanskrit Library have laid the groundwork for text digitization and preliminary morphological tagging [4], [21], [24], [25]. Treebank initiatives by the University of Zurich, INRIA and the University of Hyderabad have further enriched syntactic annotations [3], [26], [27]. Tagging tools like SanskritTagger [1] and deep learning approaches [6], [14], [15] are increasingly being adopted to enhance POS tagging, syntactic parsing, and semantic role labeling.

This paper aims to consolidate the growing body of research on computational approaches to the Rigveda, providing a comprehensive review of current methodologies, challenges, and future opportunities. The discussion is organized around

DOI: https://doi.org/10.71058/jodac.v9i6007

core NLP tasks: corpus development, morphological analysis, syntactic and semantic parsing, and high-level applications such as semantic search and question-answering systems [18], [19]. As a result this paper paves the way for new explorations in digital philology, comparative linguistics, and cultural heritage preservation.

II. Historical and Linguistic Background

Vedic Sanskrit, the language of the Rigveda, predates Classical Sanskrit and serves as the earliest attested form of the Indo-Aryan languages. It is significantly different from its classical counterpart in phonology, morphology, syntax, and lexicon. Much like what Homeric Greek is to Classical Greek. These differences are crucial for computational linguists attempting to apply modern Natural Language Processing (NLP) techniques to ancient texts.

Phonologically, Vedic Sanskrit is marked by the use of pitch accent, a feature (core to Proto-Indo-European) that has disappeared in later stages of the language. The tonal variations—uda ̄tta (high), anuda ̄tta (low), and svarita (falling)—play a semantic role and influence grammatical distinctions. From an NLP perspective, modeling such prosodic features is challenging due to limited acoustic data and the absence of explicit tonal markings in written transcriptions [4], [9]. Though some have managed to incorporate this feature into their corpus in a very efficient manner through the careful use of diacritics [21].

Morphologically, Vedic Sanskrit exhibits a complex system of nominal and verbal inflection, with more extensive case distinctions and verb forms than found in Classical Sanskrit. The language includes archaic dual and plural forms, rare verb moods, and compound constructions that are heavily context-dependent [1], [12]. These factors complicate tasks like lemmatization and morphological disambiguation, requiring specialized rule-based or hybrid approaches [6], [14]. Essentially it has a lot more prepositional dependence and variety than it's classical counterpart.

Syntactically, the Rigveda features a relatively free word order due to its inflectional richness, though Subject-Object-Verb (SOV) is dominant. Deviations from standard syntactic patterns are frequent, especially given the metrical and poetic constraints of the text. Enjambment, ellipsis, and deliberate syntactic inversions are common rhetorical devices, complicating syntactic parsing [5], [13]. Moreover, the absence of punctuation in the original manuscripts makes sentence boundary detection an open research challenge. Although the metrical regularity largely solves this issue, there are cases where this fails due to the challenge of reconstructing larger syllable clusters lost during transmission [21].

Semantically, the Rigveda is thought to contain deeply symbolic and philosophical content. Words and phrases often have multiple layers of meaning depending on the theological, ritualistic, and cosmological context. For example, the term "Agni" (fire) can refer to the physical element, the deity, or the concept of energy or transformation. This is further exacerbated when one follows existing works where word meanings are vague at best and conflicting at worst, often within the same piece of work [22]. Understanding such polysemy is essential for semantic analysis and requires ontological modeling [11], [19].

The historical transmission of the Rigveda also poses unique challenges. It was orally transmitted for centuries with meticulous attention to pronunciation, intonation, and recitation patterns. This oral tradition preserves certain phonetic features not easily captured in writing. This makes Vedic Sanskrit the only ancient language whose pronunciation we know for certain. Recent efforts to integrate prosodic information from recitation manuals and traditions are opening new avenues for computational modeling [7], [8].

From a historical perspective, the Rigveda provides evidence for early Indo-European linguistic features. Comparative philology has long used Rigvedic data to trace linguistic evolution across the Indo-European family. Computational methods now enable more precise modeling linguistic drift, lexical borrowing, and syntactic change using annotated Rigvedic corpora [16], [20], [23].

Overall, any NLP system targeting the Rigveda must account for its linguistic uniqueness. Standard pre-trained language models or parsers trained on modern data will most likely fail without proper adaptation. Domain-specific resources, such as the Digital Corpus of Sanskrit (DCS), Sanskrit Heritage Platform, The Metrically Restored Rigveda and Vedic Treebanks, form the backbone of computational studies [3], [4], [10], [21], [26], [27]. Leveraging these tools while developing new datasets with custom tokenizers and algorithms tailored to Vedic language features is important for future progress.

III. Existing Computational Efforts

Computational linguistics applied to Vedic Sanskrit and the Rigveda has evolved substantially in the past two decades. Despite being a low-resource language, the integration of ancient Sanskrit texts into the digital and computational paradigm has seen meaningful strides across three foundational areas: corpora development, morphological analysis, and syntactic-semantic parsing. This section surveys the most significant contributions to each area and their impact on the broader goal of computational Vedic philology.

A. Corpora Development

The creation of digital corpora is a prerequisite for any NLP pipeline. The Digital Corpus of Sanskrit (DCS) is one of the earliest and most influential initiatives in this direction. It provides tokenized and morphologically annotated Sanskrit texts, though its primary focus remains Classical Sanskrit [24]. The GRETIL (Go ̈ttingen Register of Electronic Texts in Indian Languages) project and TITUS (Thesaurus Indogermanischer Text- und Sprachmaterialien) have contributed digital versions of the Rigveda, albeit in transliterated forms without extensive linguistic annotation [25], [28].

The Metrically Restored Rigveda [21] is one of the few corpora dedicated to the Rigveda that systematically restores the

DOI: https://doi.org/10.71058/jodac.v9i6007

padapa¯t.ha by taking advantage of already available knowledge around Vedic Sanskrit metre. At places many syllables have also been restored by considering the metre as the constant. This corpus also incorporates the pitch accent information by identifying frequently occurring clusters of notes and assigning diacritics to represent them. Additionally, every pada is also prefixed with a preamble that identifies the mandala, the sukta, the mantra and the pada itself which makes this an ideal candidate to be used as a dataset especially when one is performing a comparative analysis within the Rigveda.

The INRIA Sanskrit Treebank, developed by Huet and colleagues, and the University of Hyderabad Vedic Treebank Project are tailored specifically for Vedic Sanskrit [26], [27]. These resources go beyond simple digitization by providing syntactic annotations and dependency trees. They allow for training and evaluating syntactic parsers and serve as gold standards for supervised learning tasks. However, the size of these corpora remains limited due to the time-intensive nature of expert annotation.

More recently, initiatives like the Sanskrit Wordnet and the UoH's Ontological Lexicon Project have begun to integrate lexical semantics into corpora by associating word senses with concepts and philosophical categories [17]. Such linked lexical data enable semantic enrichment and are vital for ontology-driven NLP tasks such as semantic search and knowledge graph construction [14], [19].

B. Part-of-Speech Tagging and Morphological Analysis

POS tagging in Vedic Sanskrit is particularly challenging due to the language's high degree of variance such as morphological ambiguity, free word order, and use of compounds. Early work in this domain was predominantly rule-based. Tools like SanskritTagger introduced stochastic models for POS tagging and achieved moderate success on Classical Sanskrit corpora [1]. However, applying the same models directly to Vedic texts yields suboptimal results due to phonological and syntactic differences [9].

Morphological analyzers like the Sanskrit Heritage Platform utilize a finite-state transducer approach and offer comprehensive morphological breakdowns [12]. However, the tool was initially designed for Classical Sanskrit and requires substantial modifications for Rigvedic data. Researchers have attempted to overcome this limitation by introducing additional layers of grammatical analysis specific to Vedic inflection patterns [14], which could work out for the later parts of the Rigveda where the language appears closer to its classical form.

Machine learning approaches, including Conditional Random Fields (CRFs), Hidden Markov Models (HMMs), and BiLSTM-based neural networks, have shown promise in improving tagging accuracy [6], [14]. Hybrid models combining statistical tagging with rule-based post-processing have yielded better results especially when trained on partially annotated Vedic corpora.

C. Syntactic and Semantic Parsing

Parsing the syntactic structure of Rigvedic Sanskrit has proven to be one of the most challenging tasks in computational linguistics due to metrical constraints and stylistic inversion. The INRIA Sanskrit Treebank has played a critical role in enabling dependency parsing using traditional algorithms as well as neural networks [26]. Work by Goyal and Huet demonstrates the feasibility of rule-enhanced dependency parsing for morphologically rich and syntactically flexible languages [6].

For semantic parsing WordNets combined with Treebanks has been explored [11] Additionally we have Multi-layer Annotation of the Rigveda [23] which is by far the most complete annotation of the Rigveda. These initial models require further tuning but indicate that domain-specific semantic frameworks are essential for meaningful progress.

Recent work is exploring neural language models and transformers. A ByT5 based model pre-trained on Sanskrit corpora showed promising applicability to Vedic syntax after fine-tuning [15]. However, these models are data-hungry and thus constrained by the limited availability of high-quality annotated data.

Despite numerous challenges, significant computational advances have laid the groundwork for deeper analysis of the Rigveda. Existing tools provide the building blocks for future NLP systems tailored to ancient texts. Nevertheless, comprehensive progress will require more annotated resources, deeper semantic models, and interdisciplinary collaboration between linguists, computer scientists, and Sanskrit scholars.

IV. NLP Challenges in Rigveda

Natural Language Processing (NLP) applied to the Rigveda encounters a unique set of challenges arising from the text's linguistic, poetic, and philosophical complexity. These challenges span multiple levels of analysis—phonological, morphological, syntactic, and semantic—and are exacerbated by resource limitations and the divergence of Vedic Sanskrit from contemporary language models. This section identifies and elaborates on the most prominent obstacles.

Ambiguity in Compound Constructions

The Rigveda features an extensive use of compound words (sama¯sa), which can encode complex syntactic and semantic relationships within a single token. These compounds may be descriptive (bahuvr¯ıhi), determinative (tatpurus.a), or coordinative (dvandva), and are often elliptical, omitting critical syntactic elements [1], [2], [6]. Disambiguating the internal structure and semantic function of such compounds is non-trivial, especially since the same compound can convey different meanings depending on the philosophical or ritual context [14]. Current morphological analyzers often misclassify these structures due to a lack of contextual information and annotated training data tailored to Rigvedic usage [12].

DOI: https://doi.org/10.71058/jodac.v9i6007

Poetic Syntax and Non-linear Structure

The poetic nature of the Rigveda introduces syntactic inversions, metrical rearrangements, and rhetorical devices that deviate significantly from the canonical SOV order [5], [13]. Enjambment—where a phrase extends beyond the verse line—further complicates parsing. These non-linearities disrupt sentence boundary detection and syntactic parsing tasks, rendering standard dependency parsers ineffective unless trained on Vedic-specific syntactic annotations [10], [23]. Moreover, metrical constraints influence word choice and order, meaning that syntactic structures must often be interpreted in light of prosodic and stylistic factors [4].

Scarcity of Annotated Data

High-quality annotated corpora for Vedic Sanskrit remain limited in size and scope. While efforts such as the UZH Vedic Treebank, the INRIA Sanskrit Treebank, the UoH Vedic Treebank and the Multi-layer Annotation of the Rigveda have made significant contributions [3], [23], [26], [27], the volume of annotated data is insufficient to train modern data-hungry models like transformers [15]. This data scarcity also affects evaluation: without reliable benchmarks, it is difficult to measure progress or compare models. Semi-supervised learning and transfer learning from related domains offer partial solutions, but the linguistic divergence between Classical and Vedic Sanskrit limits their effectiveness [9].

Semantic Ambiguity and Contextual Interpretation

The Rigveda's semantic structure is layered and context-dependent. Hymns often employ metaphor, allegory, and symbolism, with many terms bearing multiple meanings. For example, terms like "soma" may denote a ritual substance, a deity, or an abstract principle [11]. Another similar example being "gravan" [22]. Capturing such polysemy requires sophisticated semantic models grounded in domain-specific ontologies [19]. Traditional semantic role labeling frameworks are ill-suited to this task without significant adaptation, as they fail to account for the fluid conceptual mappings typical of Vedic philosophy.

Preservation of Philosophical and Cultural Context

Many NLP models prioritize surface-level syntactic and semantic analysis, often neglecting the rich philosophical and cultural subtext embedded in the Rigveda. A literal translation or extraction may miss deeper esoteric meanings unless supplemented with commentarial traditions or philosophical metadata [18]. Capturing this nuance computationally remains a major research gap. Some researchers propose integrating NLP with knowledge representation frameworks and expert-curated ontologies to bridge this gap [11], [19].

Challenges from Oral Transmission

Unlike modern written texts, the Rigveda was preserved through oral tradition, with intricate recitation patterns such as padapa¯t.ha and krama-pa¯t.ha serving to encode syntactic and phonetic information [7], [8]. These patterns are often not represented in digital text, leading to a loss of prosodic cues that are semantically relevant. Modeling these oral features computationally would require multimodal NLP approaches that integrate textual, phonetic, and rhythmic data—a task still in its infancy.

The unique structure and content of the Rigveda challenge every stage of the NLP pipeline—from tokenization to semantic interpretation. Addressing these requires not only advanced machine learning but also the integration of linguistic scholarship, digital philology, and computational creativity. Building richer annotated corpora, developing context-aware models, and designing tools specific to the Vedic domain are essential steps toward effective NLP for the Rigveda.

V. Applications and Future Directions

The unique linguistic, philosophical, and cultural features and the historical significance of the Rigveda open up a wide array of applications for NLP, ranging from academic research and cultural preservation to great strides in the field of Indo-European linguistics. As computational tools become more sophisticated, their applicability to ancient texts like the Rigveda becomes increasingly viable. This section outlines promising applications and the future research directions that can bridge existing gaps and expand the impact of NLP in Vedic studies.

Semantic Search and Information Retrieval

One of the most immediate applications of NLP for the Rigveda is the development of semantic search engines. These systems can enable scholars and students to query Vedic hymns using natural language questions, leveraging underlying syntactic and semantic structures [11], [19]. Such systems must move beyond keyword matching and incorporate contextual disambiguation, synonym recognition, and phrase structure alignment to retrieve relevant verses. This is especially important due to the limited corpora of the Rigveda hence understanding the semantic meaning of words in a reproducible way could help overcome this issue i.e. scholars are no longer constrained by the limited number of occurrences of a given word in the Rigveda. Integrating these capabilities requires semantic role labeling, word sense disambiguation, and ontology-backed search algorithms.

Vedic Question Answering (QA) Systems

Building intelligent QA systems capable of interpreting and answering queries based on Rigvedic content is a significant frontier [18]. These systems can support both factoid and explanatory questions, such as identifying deities associated with specific hymns or interpreting the symbolism in a verse. Training such models would require annotated QA datasets [23], domain-specific embeddings, and attention-based architectures like transformers [15]. Incorporating philosophical commentaries as auxiliary input may also improve contextual understanding [13].

DOI: https://doi.org/10.71058/jodac.v9i6007

Digital Preservation and Cultural Heritage

NLP has the potential to enhance the digital preservation of Vedic knowledge. Annotated corpora, prosody-aware recitations, and linked open data formats can create interactive platforms for studying the Rigveda. These tools not only aid researchers but also allow a broader audience to access and understand ancient wisdom [7], [8], [19]. Collaborations with cultural and linguistic archives, such as the Sanskrit Library and GRETIL, are essential to unify data formats and maintain historical accuracy [4].

Text Alignment and Translation Aids

Automatic alignment of Rigvedic verses with their corresponding translations and commentaries can assist scholars in comparative linguistic studies. NLP tools can help identify phrase-level alignments and semantic equivalences, enabling more accurate translations and comparative philology research [16], [20]. Advanced machine translation models, adapted for low-resource, morphologically-rich languages, can further assist in producing reliable translations, especially when guided by alignment metadata.

Cross-linguistic and Historical Linguistics Research

The Rigveda, as an ancient Indo-European text, is a goldmine for comparative linguistics. NLP models trained on Rigvedic data can be employed to study language drift, lexical borrowing, and syntactic change across Indo-European languages [16], [20]. These models can also aid in reconstructing proto-languages and testing hypotheses in historical linguistics by generating synthetic corpora and aligning parallel structures across languages.

Cognitive and Philosophical Modeling

Beyond linguistic tasks, NLP may assist in modeling Vedic cognitive and philosophical structures. Conceptual metaphor theory and frame semantics, for example, can be employed to uncover embedded ontologies and worldviews encoded in the hymns [11], [14], [19]. Tools that visualize metaphoric networks, deity relationships, and ritualistic associations can serve educational, theological, and cultural purposes.

Future Directions

To realize these applications, several future research avenues must be pursued:

- Development of large-scale, high-quality annotated corpora specifically for Vedic Sanskrit [3], [21], [23].
- Integration of oral tradition features such as recitation metrics and tonal markers into digital models [21].
- Creation of Vedic-specific word embeddings and transformer models [14], [15].
- Multilingual and cross-cultural ontology mapping to relate Rigvedic concepts to modern knowledge frameworks [19], [20].
- Hybrid NLP approaches combining rule-based Vedic grammar with machine learning algorithms [1], [6].

VI. Conclusion

NLP techniques, when adapted thoughtfully, can unlock unprecedented access to ancient wisdom. The Rigveda, as a linguistic and cultural artifact, offers a rich testbed for advancing computational linguistics and the field of Indo-European studies. Future research must focus on creating richer annotated resources, developing domain-specific models, and fostering interdisciplinary collaborations.

References

[1] Hellwig, O. (2009). *SanskritTagger: A Stochastic Lexical and POS Tagger for Sanskrit*. In: Huet, G., Kulkarni, A., & Scharf, P. (Eds.), *Sanskrit Computational Linguistics 2007/2008* (LNCS, Vol. 5402, pp. 266–277). Springer. [Springer]

[2] Krishnan, S., & Kulkarni, A. (2019). *Sanskrit Segmentation Revisited*. In *Proceedings of the 16th International Conference on Natural Language Processing (ICON 2019)* (pp. 105–114). NLP Association of India. [PDF]

[3] Hellwig, O., Scarlata, S., Ackermann, E., & Widmer, P. (2020). The Treebank of Vedic Sanskrit. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)* (pp. 5137–5146). ELRA. [PDF]

[4] Scharf, P. M., & Hyman, M. D. (2023). *The Sanskrit Library: A Digital Infrastructure for Sanskrit Texts, Linguistics, and Lexical Data*. Proceedings of the 12th Workshop on South and Southeast Asian Natural Language Processing (SANLP), pp. 12–20. [Project]

[5] Hellwig, O., Nehrdich, D., & Sellmer, S. (2023). *Data-driven Dependency Parsing of Vedic Sanskrit*. *Language Resources and Evaluation*, 57, 1173–1206. [PDF]

[6] Gupta, A., Krishna, A., Goyal, P., & Hellwig, O. (2020). *Evaluating Neural Morphological Taggers for Sanskrit*. arXiv preprint. [PDF]

[7] Aralikatte, R., Gantayat, N., Panwar, N., Sankaran, A., & Mani, S. (2018). *Sanskrit Sandhi Splitting using seq2(seq)2^*. In *Proceedings of EMNLP 2018* (short papers), pp. 487–492. [PDF]

[8] Jha, G. N. (Ed.). (2010). *Sanskrit Computational Linguistics: 4th International Symposium, New Delhi, December 10–12, 2010, Proceedings*. Lecture Notes in Computer Science, Vol. 6465. Springer. [Springer]

[9] Biagetti, E., Hellwig, O., Scarlata, S., Ackermann, E., & Widmer, P. (2021). *Evaluating Syntactic Annotation of Ancient Languages: Lessons from the Vedic Treebank*. *Old World: Journal of Ancient Africa and Eurasia*, 1(1), 1–32. [PDF]

[10] Sandhan, J., Agarwal, A., Behera, L., Sandhan, T., & Goyal, P. (2023). *SanskritShala: A Neural Sanskrit NLP Toolkit with Web-Based Interface for Pedagogical and Annotation Purposes*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023 Demo Track)*, pp. 103–112. [PDF]

[11] Brigada Villa, L., Biagetti, E., Ginevra, R., & Zanchi, C. (2023). *Combining WordNets with Treebanks to Study Idiomatic Language: A Pilot Study on Rigvedic Formulas*. In *Proceedings of the 12th Global Wordnet Conference (GWC 2023)*, pp. 102–113. [PDF]

[12] Huet, G. (2005). *A Functional Toolkit for Morphological and Phonological Processing: Application to a Sanskrit Tagger*. Journal of Functional Programming, 15(4), 573–614. [PDF]

[13] Krishna, A., Satuluri, P., Sharma, S., Kumar, A., & Goyal, P. (2016). Compound Type Identification in Sanskrit: What Roles do the Corpus and Grammar Play? In *Proceedings of WSSANLP 2016*, pp. 1–10. [PDF]

[14] Bollineni, V., Crk, I., & Gultepe, E. (2025). *Mapping Hymns and Organizing Concepts in the Rigveda: Quantitatively Connecting the Vedic Suktas*. *NLP4DH @ NAACL* (NAACL 2025 Workshop on NLP for Digital Humanities). [PDF]

[15] Nehrdich, S., Hellwig, O., & Keutzer, K. (2024). *"One Model is All You Need: ByT5-Sanskrit, a Unified Model for Sanskrit NLP Tasks."* Accepted for publication in *Findings of EMNLP 2024*. arXiv preprint. [PDF]

[16] Schrader, S. R. & Gultepe, E. (2023). *Analyzing Indo-European Language Similarities Using Document Vectors*. *Informatics*, 10(4), 76. [PDF]

[17] Patyal, H. C. (2009). *Development of Sanskrit Lexicography*. *Bulletin of the Institute of Oriental Culture*, 48(2), 113–132. [PDF]

DOI: https://doi.org/10.71058/jodac.v9i6007

[18] Terdalkar, H. & Bhattacharya, A. (2023). *Framework for Question-Answering in Sanskrit through Automated Construction of Knowledge Graphs*. arXiv preprint. [PDF]
[19] Mondaca, F., & Rau, F. (2020). *Transforming the Cologne Digital Sanskrit Dictionaries into OntoLex-Lemon*. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)* (pp. 11–14). Paris: European Language Resources Association. [PDF]
[20] Prabhu, S. M., & Radhakrishnan, G. (2024). *Chronological Analysis of Rigvedic Mandalas using Social Networks*. arXiv preprint. [PDF]
[21] Slocum, J., & Thomson, K. (2007). *The Rigveda: Metrically Restored Text*. University of Texas Linguistics Research Center. [Project]
[22] Thomson, K. (2001). The meaning and language of the Rigveda: Rigvedic gra¯van as a test case. *Journal of Indo-European Studies*, 29(3–4), 295–349. [PDF]
[23] Hellwig, O., Hettrich, H., Modi, A., & Pinkal, M. (2018). Multi-layer annotation of the Rigveda. In *Proceedings of LREC 2018*. [PDF]
[24] Hellwig, O. (2010–2024). *The Digital Corpus of Sanskrit (DCS)*. A morphologically and lexically tagged Sanskrit corpus maintained by the Sanskrit Heritage Platform. [Project]
[25] Gru¨nendahl, R., & Sub-uni Goettingen (n.d.). *GRETIL – Go¨ttingen Register of Electronic Texts in Indian Languages.* An open repository of TEI-encoded Sanskrit and other Indic texts. [Project]
[26] Huet, G. (since 2009). *Sanskrit Heritage Engine & INRIA Sanskrit Treebank*. INRIA, Paris-Rocquencourt. [Project]
[27] Kulkarni, A., Satuluri, P., Panchal, S., Maity, M., & Malvade, A. (2020). *Dependency Relations for Sanskrit Parsing and Treebank Development*. In *19th International Workshop on Treebanks and Linguistic Theories (TLT 2020)*. Presented by the University of Hyderabad's Sanskrit Studies department. [PDF]
[28] Gippert, J. (1999–present). *TITUS: Thesaurus of Indo-European Text and Language Materials* [Computer database]. Universita¨t Frankfurt. [Project]