Building Digital Resources for Dharmaśāstra Texts: A Concept Mining System for Indian Knowledge System

Arooshi Nigam[1], Subhash Chandra[2]

[1]Research Scholar, Department of Sanskrit, University of Delhi, Delhi, India
anigam@sanskrit.du.ac.in
[2]Professor, Department of Sanskrit, University of Delhi, Delhi, India
schandra@sanskrit.du.ac.in

Abstract

Dharmaśāstra (DS) is a corpus of ancient Sanskrit written tradition that provides normative guidelines for ethical behavior, social norms, and legal codes, as well as stages of life, social classes, and is a foundational component of the Indian Knowledge System (IKS). The paper presents the development of digital resources for DS texts. Addressing the challenges posed by the classical Sanskrit language and the limited availability of standardized digital corpora, this study presents techniques for corpus creation, annotation, and computational processing suited to these texts. Leveraging advancements in Human Language Technologies (HLT), the research proposes scalable methods to enable effective digitization and analysis of human language. The work highlights the creation of interoperable digital archives and annotated datasets designed to support scholarly research, pedagogy, and cross-lingual applications in IKS studies and digital humanities applications. By situating these methodologies within the broader context of Indian linguistic diversity and digital humanities, the paper demonstrates how advanced digitization practices can foster wider accessibility, cultural preservation, and renewed scholarly engagement with DS texts.

Keywords: Dharmaśāstra (DS), Digitization, Preservation, Indian Knowledge Systems (IKS), Human Language Technologies (HLT), Natural Language Processing (NLP)

1. Introduction

Indian Knowledge Systems (IKS) encompasses a vast and body of knowledge developed over centuries, rooted in ancient Indian philosophical, scientific, and cultural traditions. These systems are characterized by their holistic approach, addressing diverse human concerns including ethics, law, governance, spirituality, and material well-being.

IKS reflects the depth and diversity of India's intellectual traditions, evolved over thousands of years. It offers a distinctive way of understanding the world, epistemologies, and methods of inquiry that were central to Indian civilisation. This system draws together a wide range of disciplines, blending philosophy, language, politics, law, astronomy, mathematics, medicine, and environmental studies. In IKS, different fields often come together, allowing spiritual and scientific ideas to interact and enrich each other, creating a broad and unified vision of knowledge. Rooted in classical sources like the *Vedas*, *Vedāṅgas*, *Upaniṣads*, *Purāṇa*s, *smṛtis*, and *śāstras*, IKS has always aimed to be practical, shaping not just abstract ideas but real-world practices in governance, economy, education, and society.

From the profound speculations of the *Upaniṣads* to the logical precision of *nyāya* and *vaiśeṣika*, the medical expertise of Ayurveda and the *Caraka* and *Suśruta Saṃhitā*s, the mathematical innovations of the *Śulvasūtras*, and celebrated treatises like *Arthaśāstra*, *Nāradasmṛti*, *Śukranīti*, and *Natyaśāstra*, IKS stands as both a storehouse of ancient wisdom and a living tradition of ongoing inquiry and creativity.

Within the key framework of IKS, the *Dharmaśāstra* (DS) texts hold a primal position. Composed in classical Sanskrit, these compendiums, codify the principles of dharma (righteousness), social conduct, law, and ethics. DS have provided practical guidance to individuals and prompted the societal behavior, representing the right way of living according to *dharma* (Horsch, 2004), forming a foundational pillar of Indian intellectual and cultural heritage. These texts outline the tenets of *dharma* (righteousness), *artha* (governance and wealth), and *vyavahāra* (legal procedures), reflecting India's sophisticated legal-philosophical systems.

Despite their profound significance, DS texts face challenges in accessibility and scholarship due to their complexity, language precision, and the fragmented state of available manuscripts.

Digitization of archaic texts, opens up a revolutionary possibility to preserve these texts, making them accessible to scholars, students, and practitioners worldwide. By creating standardized, searchable digital corpora and linguistic resources, digitization not only safeguards this cultural wealth from loss but also enables advanced computational study, cross-lingual analysis, and pedagogical innovation.

This paper presents a pioneering initiative to digitize and make the profound corpus of DS literature, the canonical Sanskrit texts that constitute the ethical, legal, and socio-political foundations of ancient Indian civilization accessible to the global audience.

It explores the linguistic techniques and computational approaches such as data mining, concept indexing, information extraction (IE) and information retrieval (IR) (Harter, 1975; Hjørland, 2018), essential for building comprehensive digital resources for DS texts within the broader framework of Indian Knowledge Systems, aiming to bridge the gap between ancient wisdom and contemporary digital technologies.

2. Multiplicity of Concepts in DS Texts

DS literature encompasses a remarkably wide range of topics, addressing both ancient themes, such as personal conduct, social duties, legal principles, and ethical norms,

Arooshi Nigam, Subhash Chandra, *Building Digital Resources for Dharmaśāstra Texts: A Concept Mining System for Indian Knowledge System*. In: Z. Vetulani, P. Paroubek, I. Kardava (Eds.), Human Language Technologies as a Challenge for Computer Science and Linguistics – 2025, pp. 57–61. Poznań: Adam Mickiewicz University Press, 2025. © Authors, 2025. https://doi.org/10.14746/amup.9788323245414.11. Open Access chapter, distributed under the terms of the CC licence (BY-NC-ND, https://creativecommons.org/licenses/by-nc-nd/4.0/).

which are presented through various conceptual categories like *ācāra* (conduct), *vyavahāra* (social and legal transactions), and *prayascitta* (penance or expiation), *puruṣārtha*, *varṇāśrama*, *vivāha*; as well as contemporary concerns including environment, labor, jurisprudence, mathematics, political science, military strategy, management, and social sciences. This extensive scope highlights the enduring relevance of these texts as comprehensive guides for both individual behavior and societal governance.

These texts are known for their complex classification of ideas, as different DS texts emphasize or organize concepts in unique ways according to differing regional, temporal, and doctrinal contexts. This multiplicity illustrates the dynamic evolution of dharma across centuries, where authoritative guidance accommodated varying social customs, individual roles, and emerging governance needs.

Understanding this pluralism is crucial in developing digital tools and concept mining systems capable of capturing the rich conceptual diversity embedded within DS literature, ensuring their faithful representation and nuanced analysis in modern scholarly and computational frameworks.

3. Barriers to Digitizing Classical Literature

Digitizing DS texts presents a unique set of challenges rooted in both the linguistic features of classical Sanskrit and the physical or practical challenges of ancient manuscripts.

3.1 Linguistic Complexity of Classical Sanskrit

DS texts are written in a highly formalized, intricate version of Sanskrit, marked by complex grammar, nuanced terminology, and distinctive orthographic conventions. This linguistic richness poses significant hurdles for computational processing, machine translation, and text parsing. Standard Natural Language Processing (NLP) tools, typically optimized for modern languages, often struggle to accurately handle the advanced syntactic and semantic layers of classical Sanskrit. Furthermore, the presence of commentarial literature and regional script variations increases the difficulty of standardized transcription and automatic analysis (Mittal, 2022).

3.2 Scarcity of Standardized Digital Corpora and Resources

A major obstacle in the digitization of DS literature is the lack of standardized, annotated digital corpora. Many existing texts are only available as scanned manuscripts or in limited print editions, making them inaccessible to search algorithms and advanced linguistic tools.

Another challenge is that the available resources are fragmented, due to various editions, commentaries, and script variations. It complicates the efforts for developing comprehensive, interoperable digital archives. Standardized transcription protocols and curated, peer-reviewed corpora are urgently needed to ensure both accuracy and usability for future research and digital applications.

3.3 Problems Specific to DS Manuscripts

The physical condition of DS manuscripts introduces additional digitization challenges. Variations in scripts (such as *Devanāgari*, *Grantha*, or regional forms), use of diacritical marks, and the presence of marginalia, glosses, or interpolations complicate optical character recognition (OCR) and digital text extraction.

The fragile state of many manuscripts requires careful handling and high-fidelity imaging, often necessitating manual correction and validation of digital outputs. Accurately capturing the complex layers of the primary texts along with the detailed traditional commentaries is a difficult task for both software and human editors.

4. Objectives

The objectives for carrying out this research are as follows:

- To establish standardized digitization methods for DS texts, including scanning, OCR, IE, indexing and mining suitable for Sanskrit manuscripts.
- To manually compile a comprehensive corpus by integrating manuscripts, digitized texts, and oral traditions relevant to DS literature.
- To utilize Human Language Technologies (HLT) and community-driven approaches to address resource limitations typical of classical Sanskrit texts.
- To build an easy to use, interoperable digital archives and tools that support scholarly research, academic teaching, and cultural preservation while ensuring ethical and inclusive digitization practices.

5. Computational Techniques and Tools

Developing linguistic resources is a primary yet the most, pertinent step in digitizing DS texts to enable effective computational processing and analysis. The process begins with corpus building, where raw textual data is collected from diverse sources such as ancient manuscripts, existing digitized editions, and oral traditions preserved through recitations and commentaries.

DS literature being fragmented in nature and varying across sources, thus compiling a comprehensive and balanced digital collection requires thoughtful selection and reconciliation of different versions. Once the corpus is built, it is essential to add linguistic annotations that make the texts interpretable to computers. This involves labelling morphological features and grammatical forms such as roots, suffixes, affixes, sentence structure, and meanings. Such annotation enables advanced language processing, like parsing, machine translation, and IR, by providing detailed information that supports deeper computational analysis.

Computational techniques that are majorly adopted to digitize ancient knowledge texts are:

5.1 OCR: OCRs tailored specifically for Sanskrit texts presents unique challenges due to complex ligatures, diacritical marks, and script variability. A Recently, AI-powered OCR systems trained on Sanskrit data have improved text extraction accuracy, enabling the conversion of scanned images into searchable digital content, facilitating preservation and research.

5.2 Computational parsing: It enables detailed analysis of DS texts by automatically addressing the intrinsic complexities of Sanskrit grammar, such as sandhi (word joining), varied word order, and rich morphological structure. State-of-the-art deep learning frameworks, Modern neural models and rule-based analyzers, namely ***ByT5-Sanskrit*** (Sandhan et al., 2023) and toolkits such as

SanskritShala (Nehrdich et al., 2024), enable word segmentation, lemmatization, morphological tagging, and syntactic analysis, making it possible to apply large-scale NLP techniques to classical texts.

5.3 Linguistic processing tools: Tools such as, semantic search engines, machine translation (MT) models, and pedagogical applications, offer practical ways to explore, interpret, and teach DS literature. These platforms support tasks such as semantic IR, content-based search, and automated assessment, making ancient wisdom more widely accessible to contemporary audiences. They are used in digital archives, educational platforms, and research tools, bringing classical Sanskrit literature to a wider audience and enabling in-depth computational study (Sahu and Pal, 2023).

5.4 IE/IR Tools: Extraction and retrieval refers to the set of computational methods used to automatically identify, extract, and retrieve relevant facts, entities, and other structured data from the unstructured DS texts to develop an efficient digital resource. These techniques enable users to search for and pinpoint specific content, such as laws, ethical prescriptions, or social rules; across large text corpora, transforming complex manuscripts into searchable, accessible datasets.

5.5 Concept Mining: It involves developing a robust system to identify and organize the underlying concepts embedded in DS literature. Rather than simply extracting keywords, concept mining uncovers the deeper ideas, themes, and categories present in these texts (as mentioned in section 2) and structures them for analytical use. This provides a powerful tool for mapping, visualizing, thematic analysis, and interpreting the rich conceptual landscape of classical Indian knowledge, supporting both digital research and cross-textual analysis (Feldman & Dagan, 1955).

Overall, linguistic resource creation for DS texts is a multidisciplinary effort combining philological expertise, computational linguistics, and digital humanities methodologies to transform ancient wisdom into accessible digital assets.

6. Methodology and Development of Concept Mining System

The dataset for this digital resource developmental project comprises critical texts, specifically:

1. *Āpastamba Dharmasūtra: 1373 sūtras*
2. *Gautama Dharmasūtra: 997 sūtras*
3. *Vasiṣṭha Dharmasūtra: 1,037 sūtras*
4. *Baudhāyana Dharmasūtra: 1279 sūtras*
5. *Manusmṛti: 2,699 verses*
6. *Yājñavalkyasmṛti: 1010 verses*
7. *Nāradasmṛti: 1028 verses*
8. *Śukranīti: 2567 verses*
9. *Arthaśāstra: Approximately 6000 verses*

The data sources include authoritative manuscripts, digitized editions, authentic commentaries, and translation manuscripts from recognized scholars. All data is unified in UTF-8 Unicode encoding in both Devanagari and Roman scripts, allowing flexible processing and input/output.

6.1 Research Methodology & Technical Framework

This research employs a comprehensive and integrated methodological framework designed specifically for building digital resources on DS texts:

6.1.1 Data Mining: The project extracts relevant verses, legal concepts, and structured information from classical Sanskrit DS literature, focusing on high-quality digital transcription and content curation.

6.1.2 IE: Advanced linguistic processing techniques parse complex Sanskrit texts, involving tagging, translation, and morphological analysis to identify meaningful linguistic units and relationships throughout the corpus (Hulth et al., 2001).

6.1.3 Concept Mining: The system identifies thematically significant concepts such as dharma, daṇḍa (punishment), and vivāha (marriage), enabling users to perform contextual searches that go beyond simple keyword matching, thereby facilitating deep exploration of DS ideas.

6.2 Linguistic Annotation and Preprocessing

Transliteration for digitized texts follows the IAST standard for Roman script, ensuring accuracy and interoperability across scripts. Manuscript images were digitized via OCR using Google Lens. Benchmarking studies have reported mean Character Error Rates (CER) as low as 0.146 on Sanskrit texts, corresponding to greater than 98% character-level accuracy under ideal conditions. Internal pilot tests yielded consistent text extraction accuracy in the 90-95% range for high-quality manuscript images, typically requiring minimal manual correction.

6.3 Tagging and Annotation Workflow

At this stage, conceptual annotation and tagging are conducted manually by domain experts, maximizing semantic precision for complex legal and philosophical concepts. The preliminary results show over 90% consensus on concept extraction conducted on 2000 verse pilot subset *of Manusmṛti* in the initial phase. This approach ensures that foundational semantic categories are mapped reliably prior to large-scale automation. The manual annotation pipeline thus provides the necessary ground truth for evaluating automated approaches in later project phases.

6.4 Technical and Methodological Design

The current system employs a primarily manual, unsupervised methodology for concept mining to uphold high interpretative fidelity critical to Sanskrit legal texts. Future work plans to incorporate automated linguistic preprocessing, such as lemmatization and POS tagging, to enhance both corpus usability and search functionality. As annotated data grows, hybrid and supervised models, such as neural sequence tagging pipelines and ByT5-Sanskrit for token-level analysis can be integrated to enable robust, systematic, and quantitative evaluation across the corpus.

6.5 Evaluation Metrics and Future Directions

While comprehensive, large-scale quantitative evaluation (e.g., accuracy, precision, recall) is planned for subsequent releases, pilot tests already indicate that OCR achieves above 95% accuracy for suitable manuscript images, aligning with external CER benchmarks. Measures of inter-annotator agreement are being tracked; future updates will report detailed metrics for all steps as digital linguistic preprocessing tools become part of the workflow. This

strategy ensures continual assessment and transparent reporting of technical progress throughout the project lifecycle.

6.6 Developmental Process

6.6.1 Digitization: Texts are digitized in Unicode (UTF-8) format, supporting both Devanagari and Roman script (IAST) versions to enhance accessibility and scholarly use.

6.6.2 Database Creation: A bilingual database is constructed containing original ślokas alongside English and Hindi translations, ensuring a broad reach and usability for diverse audiences.

6.6.3 Tagging and Indexing: Each verse is semantically tagged to empower retrieval through both keyword-based and concept-based queries, thereby enabling effective navigation across the vast corpus.

6.6.4 Interface Design: The front-end interface is developed using HTML, CSS, and JavaScript to support dynamic interactivity, allowing inputs in both scripts. The back-end built with Python and Flask processes queries and dynamically generates search results with accuracy and speed (Bhatia et al., 2011; Zelle, 2004). The online system features a form-based user interface designed to facilitate efficient searching and concept mining within the DS corpus. Users can submit queries via two primary mechanisms: a text area and a dropdown menu. The text area allows users to input keywords in either Devanagari (UTF-8) or Roman IAST scripts, enabling retrieval of references, translations, and exegesis. Alternatively, the dropdown menu provides quick access to pre-defined concepts by allowing users to select keywords directly and receive associated verse-level details. All user inputs are processed in real-time, yielding relevant output, such as verse references and contextual information, on the same page, thereby supporting seamless interaction and exploration of the manuscript content.

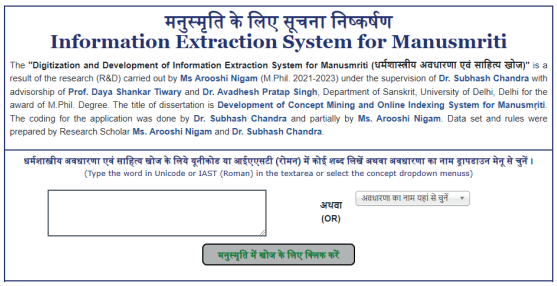

Fig. 1: System Interface Screenshot

6.7 System Architecture and Workflow

The developed system is structured into several key modules that collectively ensure seamless query handling and information delivery:

6.7.1 User Interface: Accepts user input in Devanagari or Roman script.

6.7.2 Pre-Processor: Cleans and formats the input queries to normalize variations and prepare for further processing.

6.7.3 Script Validator: Detects the script type of input and routes it accordingly for optimized processing.

Concept Validator: Matches input queries with pre-tagged concepts, ensuring semantic alignment and relevant results.

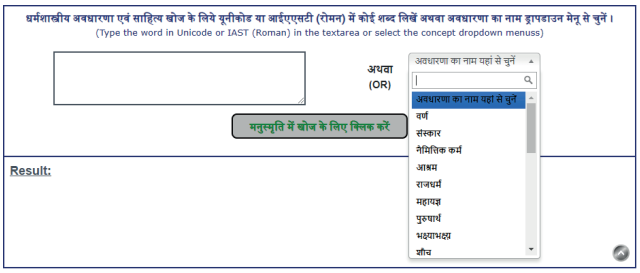

Fig. 2: Dropdown Menu

6.7.4 Information Generator: Retrieves associated ślokas, explanatory notes, and cross-references from the database.

6.7.5 Output Generator: Presents structured and user-friendly results, with options to download outputs in text or PDF formats for offline use.

This modular design guarantees that even imprecise or indirect queries yield meaningful results through the concept-based indexing and semantic tagging framework, thereby enhancing scholarly research and pedagogical applications related to DS texts.

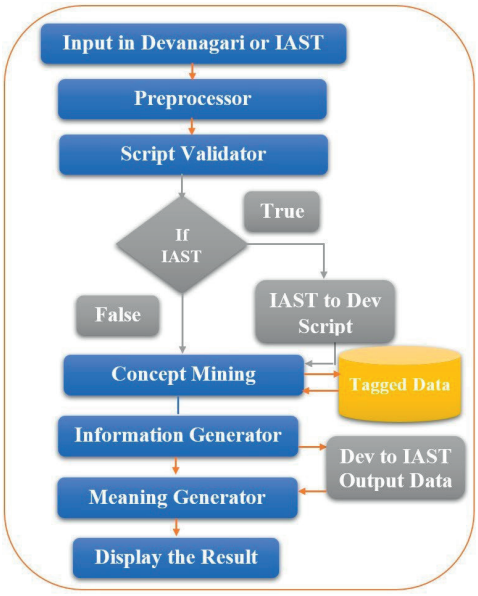

Fig. 3: The Concept Mining Pipeline Flowchart

6.8 Leveraging HLT

In the context of building digital resources for DS texts within a concept mining framework for IKS, leveraging HLT takes on unique importance. Given the low-resource and complex nature of classical Sanskrit materials, this paper adopts a multi-pronged strategy to advance digital humanities research.

6.8.1 Transfer learning is utilized to adapt state-of-the-art NLP models, initially trained on modern, resource-rich languages, to cater the particular demands of DS literature. By fine-tuning these models on smaller, domain-specific Sanskrit corpora, we can overcome the lack of large annotated datasets and significantly improve the processing accuracy for DS texts.

6.8.2 Multilingual models are incorporated to facilitate cross-lingual insights, as many ideas and concepts in DS have parallels or continuities in other Indian languages and

script traditions. These models support the recognition and mapping of conceptual frameworks across scripts and linguistic boundaries, broadening the analytical reach of the digital resource.

6.8.3 Equally important is the establishment of community-driven annotation frameworks. By involving scholars, language experts, and practitioners in the annotation process, this project ensures both the accuracy and the cultural authenticity of the digital corpus. Collaborative annotation not only raises the quality of linguistic metadata but also nurtures a shared digital heritage, making the resource dynamic, inclusive, and reflective of living traditions.

Together, these techniques allow the concept mining system to unlock the layered knowledge contained within DS texts and present it in a manner that is precise, accessible, and relevant to contemporary research and pedagogy in IKS.

7. Significance

The significance of this research lies in its ability to bridge the gap between ancient DS texts and contemporary digital humanities, making these complex classical works comprehensible and accessible to a global audience, including scholars, students, and practitioners.

By developing a sophisticated concept mining system tailored for Indian Knowledge Systems, this project not only preserves invaluable cultural and intellectual heritage but also enables advanced computational analysis, semantic search, and cross-disciplinary exploration of DS literature. This study demonstrates how computational linguistics can revitalize classical Sanskrit literature and unlock its applicability in modern discourse.

Furthermore, this research supports the democratization of ancient knowledge by providing multilingual interfaces and community-driven annotation frameworks, fostering inclusivity and collaboration among experts and learners worldwide. It offers a scalable model for digitizing other low-resource classical texts and contributes to the broader goals of cultural preservation, knowledge dissemination, and innovative scholarship in the age of digital information.

This research directly supports India's Digital Humanities movement, providing an archetype for similar digitization projects in Vedic, *smṛti*, and *śāstric* domains.

8. Conclusion and Future Directions

The research represents an important step toward the digital preservation and enhanced accessibility of DS texts within the Indigenous knowledge traditions. By integrating advanced computational techniques such as concept mining and human language technologies, the project not only unlocks the deeply layered knowledge embedded in these classical Sanskrit texts but also makes them usable for modern scholarly and educational purposes. The developed concept mining system enables contextual and semantic exploration that transcends traditional keyword searches, allowing researchers and practitioners to engage with the texts in a more meaningful and productive manner.

Looking ahead, the project will focus on expanding the scope of digitization to include other key theological texts. There are planned enhancements to incorporate tagging of additional scientific and social concepts found within these traditions. Furthermore, advancing the system's multilingual capabilities to support input and output in languages such as Punjabi, Sanskrit, Bengali, Telugu, Tamil, and Kannada will broaden access and inclusivity for a diverse user base.

Ultimately, this research aims to serve as a foundational digital humanities resource that preserves ancient wisdom while fostering ongoing interdisciplinary scholarship, cultural exchange, and educational innovation across the global IKS community.

Acknowledgment

This work is the outcome of the research project funded by the Indian Council of Social Science Research (ICSSR) under the major project scheme 2024-25 (File No.161/2024-25/ICSSR/RP/JM/OBC dated March 19, 2025). The authors are grateful for the financial support from the Indian Council of Social Science Research (ICSSR), New Delhi.

References

Bhatia, S., Tuarob, S., Mitra, P., & Giles, C. L. (2011). An algorithm search engine for software developers. In *Proceedings of the 3rd International Workshop on Search-Driven Development: Users, Infrastructure, Tools and Evaluation* (pp. 13–16).

Feldman, R., & Dagan, I. (1995). Knowledge discovery in textual databases (KDT). In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)* (pp. 112–117).

Harter, S. P. (1975). A probabilistic approach to automatic keyword indexing, part II: An algorithm for probabilistic indexing. *Journal of the American Society for Information Science, 26*(5), 280–289.

Hjørland, B. (2018). Indexing: Concepts and theory. *Knowledge Organization, 45*(7), 609–639.

Horsch, P. (2004). From creation myth to world law: The early history of Dharma. *Journal of Indian Philosophy, 32*(5–6), 430–431.*

Hulth, A., Karlgren, J., Jonsson, A., Boström, H., & Asker, L. (2001). Automatic keyword extraction using domain knowledge. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 472–482). Berlin: Springer.

Mittal, N. (2022). Guidelines and optimal methods for digitizing Sanskrit manuscripts. *International Journal for Research Trends and Innovation, 7*(12), 930–937. https://doi.org/10.5281/zenodo.XXXXX *(Add DOI if available)*

Nehrdich, S., Hellwig, O., & Keutzer, K. (2024). One model is all you need: ByT5-Sanskrit, a unified model for Sanskrit NLP tasks. *arXiv preprint* arXiv:2409.13920. https://arxiv.org/abs/2409.13920

Sahu, S. S., & Pal, S. (2023). Building a text retrieval system for the Sanskrit language: Exploring indexing, stemming, and searching issues. *Computer Speech & Language, 81,* Article 101518. https://doi.org/10.1016/j.csl.2023.101518

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: Volume 3 (System Demonstrations)* (pp. 103–112). Association for Computational Linguistics.

Zelle, J. M. (2004). *Python programming: An introduction to computer science.* Franklin, Beedle & Associates, Inc.