International Journal of Creative and Open Research in Engineering and Management
ISSN: 3108-1754 (Online)
Volume 02 Issue 11 November-2026 | Impact Factor: 3.5



Sanskrit Manuscript Understanding: A Review of OCR Correction, Linguistic Analysis and Retrieval-Augmented Question Answering

Sneha Sunil Sonawane[1], Neeta Deshpande[2]

[1]Department of Computer Engineering
Gokhale Education Society's R. H. Sapat College of Engineering, Management Studies and Research, Nashik, India
Email: snehas1329@gmail.com

[2]Department of Computer Engineering
Gokhale Education Society's R. H. Sapat College of Engineering, Management Studies and Research, Nashik, India
Email: neeta.deshpande@ges-coengg.org

How to Cite this Article:

Sonawane, S. S. (2026). Sanskrit Manuscript Understanding: A Review of OCR Correction, Linguistic Analysis and Retrieval-Augmented Question Answering. International Journal of Creative and Open Research in Engineering and Management, 02(11), 1-9. https://doi.org/10.55041/ijcope.v2i11.159

License:

This article is published under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

© The Author(s). Published by International Journal of Creative and Open Research in Engineering and Management.







https://doi.org/10.55041/ijcope.v2i11.159

Abstract

Sanskrit has a long and rich literary tradition in South Asia, with texts covering a wide range of subjects, including philosophy, religion, medicine, mathematics, astronomy, linguistics, and other areas of scholarship. A large part of this material is available in the form of manuscripts, printed books, and scanned documents. Converting these resources into digital text can support their long-term preservation and make them easier to search, study, and process using computational methods. However, producing reliable machine-readable Sanskrit text is not straightforward. The characteristics of the Devanagari script, complex word formations, sandhi, extensive grammatical variations, similar-looking characters, diacritic marks, and differences in the quality of scanned documents can all introduce errors during digitization.

Several studies have therefore focused on Sanskrit Optical Character Recognition (OCR) and the correction of errors produced by OCR systems. The availability of dedicated Sanskrit post-OCR correction datasets, together with models such as ByT5 and ByT5-Sanskrit, has contributed to improving the accuracy of digitized Sanskrit text. Research has also addressed related linguistic tasks, including word segmentation, sandhi splitting, lemmatization, and morphosyntactic analysis. In addition, Sanskrit question-answering systems based on knowledge graphs have been investigated for accessing information from Sanskrit texts. Recent studies have further examined Transformer-based OCR methods for Vedic Sanskrit, with particular attention to the recognition of Vedic accent marks.

This review examines existing research in these areas and analyzes how the individual technologies can contribute to an integrated Sanskrit manuscript understanding pipeline. The review identifies limitations in current approaches, particularly the separation between OCR correction, linguistic analysis, information retrieval and question answering. Finally, potential future



directions are discussed for developing explainable and source-grounded Sanskrit question answering systems that can operate on noisy or OCR-derived manuscript text.

Keywords: Sanskrit NLP, Sanskrit OCR, Post-OCR Correction, ByT5-Sanskrit, Sandhi Splitting, Morphological Analysis, Sanskrit Manuscripts, Question Answering, Retrieval-Augmented Generation, Explainable AI.

1. Introduction

Sanskrit has played an important role in the intellectual, literary, philosophical and scientific history of South Asia. A large collection of Sanskrit knowledge exists in the form of manuscripts, printed books and scanned documents. These resources contain information from diverse domains such as philosophy, religion, medicine, mathematics, astronomy and literature. However, although a large number of Sanskrit documents are available, Sanskrit is still considered a low-resource language from the perspective of computational linguistics because comparatively fewer high-quality machine-readable resources and tools are available.

Digitization is therefore an important step toward preserving and computationally studying Sanskrit literature. Optical Character Recognition (OCR) provides a way to convert scanned images of documents into machine-readable text. However, OCR output is not always accurate. Sanskrit presents additional challenges because of the complexity of the Devanagari script, conjunct characters, character similarity, diacritic marks, compound words and variations in printing styles. Document quality, page layout and typography can also introduce recognition errors.

Maheshwari et al. presented a major benchmark and dataset for post-OCR correction in Sanskrit. Their dataset contains approximately **218,000 sentences and 1.5 million words from 30 books**, covering multiple domains including astronomy, medicine, mathematics and philosophy. Their experiments demonstrated that pretrained sequence-to-sequence models can substantially improve OCR outputs. Their best-performing **ByT5+SLP1** configuration achieved a character error rate (CER) of 2.98% and a word error rate (WER) of 23.19%, compared with 3.89% CER and 30.23% WER for the original OCR output.

© 2026 The Author(s). Published by IJCOPE Journal. Website: https://ijcope.org/



OCR correction alone, however, is not sufficient for deep understanding of Sanskrit text. Sanskrit is morphologically rich, and words may contain grammatical information such as gender, number and case. In addition, sandhi can combine words and modify their surface forms. Consequently, downstream applications such as search, retrieval, translation and question answering benefit from linguistic processing.

Nehrdich et al. introduced **ByT5-Sanskrit**, a unified pretrained model designed for multiple Sanskrit NLP tasks. The model supports word segmentation, lemmatization, morphosyntactic tagging, OCR post-correction and dependency parsing. The authors also introduced a multitask dataset based on the Digital Corpus of Sanskrit for word segmentation, lemmatization and morphosyntactic tagging. Their work demonstrates that a Sanskrit-specific byte-level model can provide a unified basis for several downstream NLP tasks.

Another important direction is Sanskrit question answering. Terdalkar's work describes a framework for constructing domain-specific knowledge graphs from Sanskrit texts and using them for factual question answering. The framework was applied to relationships in the Mahābhārata and Rāmāyaṇa and to synonymous relationships in Bhāvaprakāśanighaṇṭu. The reported system answered approximately 50% of factual questions correctly, while the study also analyzed limitations at different stages of the pipeline.

Therefore, Sanskrit computational research has progressed across several individual components. However, there is an important opportunity to study these components together: **OCR → OCR correction → linguistic analysis → retrieval → question answering**. This review examines the development of these technologies and identifies the research gaps that motivate integrated and explainable Sanskrit manuscript understanding systems.

2. Review Methodology

This review focuses on research related to Sanskrit OCR, post-OCR correction, Sanskrit linguistic processing and question answering.

The literature considered in this review includes research papers and scholarly works addressing the following areas:

1. Sanskrit manuscript and document OCR
2. Post-OCR text correction

© 2026 The Author(s). Published by IJCOPE Journal. Website: https://ijcope.org/



3. Sanskrit-specific Transformer models
4. Word segmentation and sandhi splitting
5. Lemmatization and morphosyntactic analysis
6. Sanskrit question answering
7. Knowledge graph-based Sanskrit information extraction
8. Vedic Sanskrit OCR and accent recognition

The review particularly examines recent developments in Sanskrit computational linguistics and compares the objectives, methods and limitations of different approaches.

The literature considered in this review covers several important developments in Sanskrit language processing. These include the Sanskrit post-OCR benchmark introduced by Maheshwari et al. in 2022, the ByT5-Sanskrit model proposed by Nehrdich et al. in 2024, neural approaches for sandhi splitting, existing frameworks for Sanskrit question answering, and recent Transformer-based approaches for Vedic Sanskrit OCR. Together, these studies cover multiple stages involved in processing and understanding digitized Sanskrit documents.

Rather than discussing each study independently, the review groups the literature according to the major functional components involved in Sanskrit document processing. This organization helps examine how OCR correction can be connected with subsequent linguistic processing and question-answering tasks. It also highlights the gaps between these individual components and shows that several existing approaches still address these tasks separately.

3. Sanskrit OCR and Manuscript Digitization

Optical Character Recognition (OCR) serves as an important initial step in transforming scanned Sanskrit documents into machine-readable text. However, applying OCR to Sanskrit is challenging due to the structural complexity of the Devanagari script, conjunct characters, word boundaries, differences in fonts, and variations in the quality and layout of source documents.

Recent studies have increasingly explored deep learning methods for addressing these challenges. The reviewed literature includes CNN-based, CNN-BiLSTM, and Transformer-based techniques for Sanskrit OCR. Vedic Sanskrit presents additional difficulties because accent marks carry linguistic information and therefore need to be recognized accurately along with the text.



Tsukagoshi et al. investigated an **accent-aware OCR approach for Vedic Sanskrit using Transformer models**. Their work employs TrOCR fine-tuned on Vedic Sanskrit data to recognize printed Vedic Sanskrit in Devanagari and generate text in an ISO 15919-compliant Latin representation. The study also points to the value of matching scanned document images with available electronic texts, which can support philological studies as well as the development of future OCR training resources.

The research further indicates that several earlier Sanskrit OCR systems were primarily designed for Classical Sanskrit and did not fully consider the specific requirements of Vedic Sanskrit, especially the recognition of accent marks. This suggests that OCR methods may need to consider the differences between Sanskrit text traditions rather than treating all Sanskrit documents as the same type of OCR problem.

Another factor affecting OCR performance is the quality and diversity of the training data. Sanskrit documents may vary in terms of typeface, font size, page condition, formatting, and overall layout. Training datasets that include such variations can therefore be useful for developing OCR systems that perform reliably across different document sources.

Despite improvements in OCR models, recognition errors can still occur. For this reason, OCR can be viewed as an initial component of a broader Sanskrit document processing pipeline rather than as the final stage of text digitization.

4. Post-OCR Correction in Sanskrit

Post-OCR correction is the process of correcting errors generated by an OCR system. Instead of modifying the OCR engine itself, a correction model receives noisy OCR output and attempts to reconstruct the correct text using linguistic and contextual information.

Maheshwari et al. made an important contribution by releasing a large Sanskrit post-OCR correction benchmark. Their dataset contains approximately **218,000 manually verified sentences, 1.5 million words and material from 30 books**. The books cover different domains and demonstrate the variation present in Sanskrit documents.

The authors investigated several pretrained sequence-to-sequence models. Their strongest configuration used **ByT5 together with SLP1 phonetic encoding**. The model achieved a CER of 2.98% and WER of 23.19%, improving upon the original OCR output, which had CER of 3.89% and WER of 30.23%.



The research also analyzed common error patterns at graphemic and lexical levels. Errors may result from scan quality, typography, visually similar characters, layout differences and linguistic characteristics of Sanskrit.

This approach is significant because it demonstrates that Sanskrit OCR correction can be treated as a sequence-to-sequence generation task. Instead of making isolated character-level substitutions, a language model can consider the surrounding context and generate the most appropriate text sequence.

Nevertheless, high OCR correction accuracy does not necessarily imply linguistic accuracy. The generated output may appear to be a valid Sanskrit sequence while still containing grammatical or semantic errors. Therefore, Sanskrit-specific linguistic validation is required to assess the correctness of the generated text. Further evaluation at the linguistic and application levels is also necessary to determine the effectiveness of the correction process.

5. ByT5-Sanskrit and Unified Sanskrit NLP

The introduction of **ByT5-Sanskrit** represents an important development in Sanskrit NLP. Nehrdich et al. designed the model specifically for Sanskrit and evaluated it on several NLP tasks. These include word segmentation, lemmatization, morphosyntactic tagging, dependency parsing and OCR post-correction.

The model is based on byte-level pretrained language modeling. This is particularly useful for Sanskrit because conventional tokenization can become difficult when dealing with morphologically complex words, rare forms and spelling variations.

The authors report that ByT5-Sanskrit performs strongly on Sanskrit word segmentation and achieves state-of-the-art performance on Vedic Sanskrit dependency parsing and OCR post-correction tasks. They additionally developed a multitask dataset for joint word segmentation, lemmatization and morphosyntactic tagging using resources from the Digital Corpus of Sanskrit.

These capabilities are relevant to a manuscript understanding system because OCR correction is not an isolated task. Once text is corrected, the system can further analyze the text linguistically.

For example, the processing pipeline can be represented as:

© 2026 The Author(s). Published by IJCOPE Journal. Website: https://ijcope.org/



OCR Text → Correction → Segmentation → Lemmatization → Morphosyntactic Analysis

This allows the system to move from character-level correction toward word-level and grammatical understanding.

6. Sandhi Splitting and Sanskrit Linguistic Processing

Sandhi is a fundamental linguistic phenomenon in Sanskrit in which sound transformations occur at word boundaries. As a result, multiple words can appear as a combined surface form. Sandhi splitting, also known as *viccheda*, attempts to recover the component words.

Dave et al. formulated Sanskrit sandhi generation and splitting as sequence-to-sequence prediction tasks using neural networks. Their work emphasizes that while sandhi formation follows well-defined rules, sandhi splitting is more difficult because the splitting can be non-unique and context-dependent.

This is significant for information retrieval and question answering. A user may search for a word in its independent form while the source text contains the word in a sandhied form. Without segmentation or sandhi processing, a retrieval system may fail to identify the relevant passage.

The connection between OCR correction and sandhi processing is an important aspect of Sanskrit text analysis. OCR mistakes may appear around word boundaries, whereas Sanskrit grammar can naturally merge words through sandhi rules. As a result, an effective system must determine whether a particular form represents an OCR error or a legitimate linguistic transformation.

7. Lemmatization and Morphosyntactic Analysis

Sanskrit contains extensive inflection. The surface form of a word can vary according to grammatical properties such as gender, number and case. Verbs can similarly encode grammatical information.

Lemmatization attempts to identify the base or dictionary form of a word. **Morphosyntactic analysis** provides grammatical information such as part of speech and morphological features.



These tasks are important because information retrieval and question answering cannot always rely on exact surface-form matching. Different inflected forms may correspond to the same underlying lexical item.

ByT5-Sanskrit provides a unified framework for word segmentation, lemmatization and morphosyntactic tagging. The model's multitask approach is particularly useful because these linguistic tasks are interconnected.

For a future Sanskrit QA system, linguistic analysis can therefore be used to improve retrieval and answer interpretation.

A simplified representation is:

Surface Word → Segmentation → Lemma → Morphological Features → Semantic Retrieval

This approach can help bridge the gap between the form in which a Sanskrit word occurs in a document and the form used by a user in a query.

8. Sanskrit Question Answering and Knowledge Graphs

Sanskrit question answering (QA) has emerged as an important area of research, particularly for extracting and organizing knowledge from Sanskrit literature. Terdalkar's Sanskrit Question Answering Framework uses knowledge graphs to represent information from specific domains and provide answers to factual questions posed in Sanskrit.

The framework develops domain-specific knowledge graphs and evaluates them using examples related to human relationships described in the *Mahābhārata* and *Rāmāyaṇa*. It also considers synonym relationships obtained from *Bhāvaprakāśanighaṇṭu*, an important Ayurveda text. The reported experiments achieved approximately 50% success for factual question answering.

The question processing procedure involves several linguistic stages. Compound expressions are processed using Sanskrit sandhi and compound-splitting techniques, after which the individual components are analyzed with available Sanskrit linguistic resources. Relevant nouns and relationship terms are identified and represented as query triplets for retrieving the required information.

However, this approach mainly addresses factual question answering over structured knowledge stored in knowledge graphs. It does not specifically address question answering from scanned Sanskrit manuscripts containing OCR errors.



9. Comparative Analysis of Existing Approaches

The reviewed research can be summarized as follows:

Research Area	Representative Work	Main Technique	Major Contribution	Main Limitation
Sanskrit OCR Correction	Maheshwari et al. (2022)	ByT5 + SLP1	Large post-OCR benchmark and strong correction results	Focuses mainly on correction
Sanskrit NLP	Nehrdich et al. (2024)	ByT5-Sanskrit	Unified segmentation, lemmatization, morphology and OCR correction	Does not provide a complete manuscript-to-QA framework
Sandhi Splitting	Dave et al. (2021)	Neural Seq2Seq	Data-driven sandhi generation and splitting	Splitting remains context-dependent
Sanskrit QA	Terdalkar Graph	Knowledge	Domain-specific factual QA	Limited domain and structured-question assumptions
Vedic Sanskrit OCR	Tsukagoshi et al. (2025)	TrOCR/Transformer	Accent-aware Vedic OCR	Focuses on Vedic OCR rather than full downstream QA



The reviewed studies indicate considerable development across different areas of Sanskrit language processing. The post-OCR benchmark introduced in 2022 provides a useful basis for evaluating OCR correction methods, while ByT5-Sanskrit demonstrates the application of Sanskrit-focused language models to multiple linguistic tasks. Neural approaches for sandhi splitting further support automated linguistic analysis, and knowledge-graph-based question answering shows the potential for retrieving structured information from Sanskrit texts. In addition, recent research on Vedic Sanskrit OCR highlights the need for specialized models that can handle variations within different forms and textual traditions of Sanskrit.

10. Identified Research Gaps

The literature review reveals several important research gaps.

10.1 Lack of an End-to-End Pipeline

Most reviewed studies focus on a particular task such as OCR correction, linguistic analysis or question answering. There is comparatively less emphasis on a complete pipeline that begins with a scanned Sanskrit document and ends with a source-grounded answer.

10.2 Limited Integration of OCR Correction and Linguistic Analysis

OCR correction improves the text, but subsequent linguistic tasks can introduce additional challenges. An integrated system should consider segmentation, morphology and grammatical information while processing corrected text.

10.3 Need for Sanskrit-Aware Retrieval

Traditional retrieval methods may depend heavily on exact word matching. Sandhi, inflection and OCR errors can reduce retrieval effectiveness. Sanskrit-aware retrieval should account for linguistic variants and corrected forms.

10.4 Limited Explainability

A correction model may change an OCR-generated word, but users may not know why the change was made. For academic and manuscript research, it is valuable to show the original OCR form, corrected form and evidence supporting the correction.

10.5 Limited Source-Grounded Question Answering

A QA system should not only generate an answer but also provide the Sanskrit passage from which the answer was derived. This is especially important for scholarly applications where users need to verify the original source.

© 2026 The Author(s). Published by IJCOPE Journal. Website: https://ijcope.org/



10.6 Vedic and Classical Sanskrit Differences

Recent Vedic OCR research demonstrates that Sanskrit variants can have different computational requirements. Vedic accent marks are an example of information that may not be handled adequately by systems designed primarily for Classical Sanskrit.

11. Future Research Directions

Future research can focus on integrating Sanskrit OCR correction with linguistic analysis and question-answering systems. Sanskrit-specific models can be further explored for improved text correction, segmentation and morphological analysis. Retrieval-based approaches can also be used to provide source-grounded answers from Sanskrit texts.

12. Conclusion

This review examined major developments in Sanskrit OCR, post-OCR correction, linguistic analysis and question answering. Existing research demonstrates significant progress in each of these areas. Maheshwari et al. introduced a large Sanskrit post-OCR benchmark and demonstrated the effectiveness of byte-level sequence-to-sequence models. Nehrdich et al. further demonstrated the potential of a Sanskrit-specific unified model, ByT5-Sanskrit, for segmentation, lemmatization, morphosyntactic analysis, OCR correction and other tasks. Neural approaches have also been applied to sandhi splitting, while knowledge-graph-based frameworks have demonstrated the feasibility of Sanskrit question answering. Recent Transformer-based research on Vedic Sanskrit OCR highlights the importance of specialized handling of accent marks and domain-specific characteristics.

Despite these advances, the literature indicates that Sanskrit OCR correction, linguistic processing, retrieval and question answering are often treated as separate problems. An integrated approach that connects these stages could provide a more useful pathway for Sanskrit manuscript understanding.

Future research can therefore focus on **explainable OCR correction, Sanskrit-aware linguistic processing, retrieval from curated Sanskrit corpora and source-grounded question answering**. Such systems have the potential to



support digital preservation, Sanskrit research, education and wider access to the knowledge contained in historical Sanskrit literature.

References

1. Maheshwari, A., Singh, N., Krishna, A., & Ramakrishnan, G. (2022). **A Benchmark and Dataset for Post-OCR Text Correction in Sanskrit.** Findings of the Association for Computational Linguistics: EMNLP 2022, 6258–6265.
2. Nehrdich, S., Hellwig, O., & Keutzer, K. (2024). **One Model is All You Need: ByT5-Sanskrit, a Unified Model for Sanskrit NLP Tasks.** Findings of the Association for Computational Linguistics: EMNLP 2024, 13742–13751.
3. Dave, S., Singh, A. K., Prathosh, A. P., & Lall, B. (2021). **Neural Compound-Word (Sandhi) Generation and Splitting in Sanskrit Language.** Proceedings of CODS-COMAD 2021.
4. Terdalkar, A., & Bhattacharya, P. **Framework for Question-Answering in Sanskrit through Automated Construction of Knowledge Graphs.**
5. Terdalkar, A. (2023). **Sanskrit Knowledge-based Systems: Annotation and Computational Tools.** PhD Thesis.
6. Tsukagoshi, Y., Kuroiwa, R., & Ohmukai, I. (2025). **Towards Accent-Aware Vedic Sanskrit Optical Character Recognition Based on Transformer Models.** Computational Sanskrit and Digital Humanities, World Sanskrit Conference 2025, 70–80.
7. Krishna, S. et al. (2018). **Post-editing / OCR correction approaches for Sanskrit text.** Discussed in the reviewed Vedic Sanskrit OCR literature.
8. Hellwig, O. & Nehrdich, S. (2018). **Sanskrit linguistic analysis and compound/sandhi processing resources.** Referenced in the Sanskrit QA framework.

© 2026 The Author(s). Published by IJCOPE Journal. Website: https://ijcope.org/