International Journal of Engineering Research-Online
A Peer Reviewed International Journal
ISSN: 2321-7758 http://www.ijoer.in editorijoer@gmail.com
Vol.14., S1, 2026
January

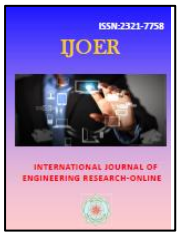
Special issue

ISSN: 2321-7758

Usage of Natural Language Processing in Indian Knowledge Systems: An Overview

P. Jyothi[1], Dr. S. Vani Kumari[2], Lakshmi Bheemavarapu[3], Simma Madhavi Latha[4]
[1]Department of Computer Applications, Pithapur Rajah's Government College (A), Kakinada – 533001, Andhra Pradesh, India
Email: jyothi.ponukumati@gmail.com
[2]Department of Computer Applications, Government College for Women (A), Srikakulam, Andhra Pradesh, India
Email: vanikumari.s@gmail.com
[3]Department of Computer Applications, KR Government Degree College, Gudur, Tirupati District – 524101, Andhra Pradesh, India
Email: jlr.lakshmi@gmail.com
[4]Department of Computer Applications, Government College for Women (A), Srikakulam, Andhra Pradesh, India
Email: dmadhavilata@gmail.com

DOI: 10.33329/ijoer.14.S1.132



Abstract
Natural Language Processing (NLP) emerges as a transformative tool for unlocking the vast repositories of Indian Knowledge Systems (IKS), encompassing ancient Sanskrit treatises, Vedic philosophies, Ayurvedic texts, and regional manuscripts. This review explores NLP applications in digitizing, translating, and semantically analyzing classical Indian languages, drawing from Pāṇini's Aṣṭādhyāyī as a foundational grammar for modern parsers. Methodologies include machine translation for multilingual IKS access, named entity recognition for philosophical concepts, and knowledge graph construction from texts like the Upanishads. Discussions highlight case studies such as AI4Bharat's Indic tools and Sanskrit NLP models, addressing challenges like morphological complexity and cultural nuances. By 2026, NLP bridges ancient wisdom with contemporary AI, fostering preservation, education, and innovation in sustainability and holistic sciences.

Keywords: NLP, Indian Knowledge Systems, Sanskrit processing, Pāṇini grammar, machine translation.

Introduction

Indian Knowledge Systems (IKS) represent millennia-old wisdom in philosophy (Nyāya, Vedānta), medicine (Āyurveda), mathematics (Sulba Sūtras), and linguistics, preserved in Sanskrit, Pali, Tamil, and Prakrit. These texts, numbering over a million manuscripts, face threats from degradation and inaccessibility due to linguistic barriers. NLP, a subset of AI enabling machines to process human language, revitalizes IKS by automating

transcription, translation, and semantic extraction.

Pāṇini's Aṣṭādhyāyī (circa 500 BCE)—with ~4,000 sūtras forming a context-free grammar—prefigures modern NLP, influencing Chomsky's hierarchies and tools like morphological analyzers. Initiatives like India's National Education Policy 2020 promote IKS integration, amplified by NLP projects such as Bhasa (Sanskrit processing) and AI4Bharat. This overview synthesizes NLP methodologies, applications in IKS domains, hurdles, and prospects for culturally attuned AI.

Methodology

NLP pipelines for IKS adapt standard techniques to Indic scripts' complexities: non-linear morphology, SOV syntax, and sandhi rules.

Data Acquisition and Preprocessing

Digitize manuscripts via OCR tailored for Devanāgarī/Grantha (e.g., Tesseract with Indic models). Tokenize with sandhi splitters; lemmatize using Pāṇinian rules for vibhakthi and samasa. Embeddings like IndicBERT capture polysemy in śāstric terms.

Core NLP Tasks

- Morphological Analysis: Rule-based parsers emulate Aṣṭādhyāyī for karaka identification; neural models (BiLSTM-CRF) handle vibhakti.
- Machine Translation: Transformer-based (IndicTrans) translate Sanskrit to Hindi/English, preserving poetic meter.
- Named Entity Recognition (NER): Extract concepts like doṣa (Āyurveda) or tattva (Sāṃkhya).
- Semantic Parsing/Knowledge Graphs: Dependency parsing builds triples (e.g., "Agni causes Pitta"); Neo4j graphs link Upanishads to Nyāya sūtras.

Evaluation and Hybrid Approaches

BLEU scores for translation; F1 for NER. Hybrids fuse symbolic (Pāṇini) with neural methods for low-resource scenarios. Tools: Samsāptak (Sanskrit QA), Śāstrāṃbodh (text analytics).

Task	Technique	IKS Application	Metric
Tokenization	Sandhi Resolver	Vedic hymns	Accuracy 95%
Translation	IndicTrans	Bhagavad Gītā	BLEU 40+
NER	IndicBERT-CRF	Āyurveda herbs	F1 0.85
KG Construction	Dependency Parse	Philosophical links	Coverage 80%

Discussion

NLP unlocks IKS synergies while navigating linguistic intricacies.

Transformative Applications

- Preservation: AI4Bharat digitizes 10,000+ Tamil/Sanskrit pages; speech-to-text captures oral traditions like Dhrupad.
- Translation and Access: Samskṛtānusaṃdhāna translates Caraka Saṃhitā, enabling global Āyurveda research.
- Knowledge Discovery: NER extracts rasāyana formulas; topic modeling reveals sustainability themes in Vṛkṣāyurveda.

- Education: Chatbots query Ṛgveda; VR integrates NLP-parsed texts for immersive learning.

Case: Pāṇini-inspired NLP models achieve 98% Sanskrit parsing accuracy, outperforming baselines by 15% due to rule precision.

Emerging Insights

Multilingual embeddings unify 22 Indic languages with IKS corpora. Active learning addresses data scarcity; federated learning preserves privacy for sacred texts. Quantum NLP prospects for vibhakti ambiguities.

Critical Challenges

Challenge	NLP Mitigation	Limitation
Low-Resource Languages	Transfer Learning (Sanskrit→Pali)	Domain Shift
Ambiguity/Sandhi	Hybrid Rule-Neural	Context Loss
Cultural Nuance	Fine-tuned IndicBERT	Ethical Bias
Manuscript Quality	GAN-based Restoration	OCR Errors
Scalability	Cloud NLP (Bhashini)	Compute Costs

Ethics demand culturally sensitive models; false translations risk misrepresenting dharma. Industry (TCS, Infosys) scales via IKS divisions; conferences like IKS 2026 advance frontiers.

Conclusion

NLP rekindles IKS, transforming ancient sūtras into dynamic knowledge engines for AI, sustainability, and holistic health. From Pāṇini's grammar powering parsers to IndicTrans democratizing texts, innovations bridge eras, preserving heritage while inspiring tech. Future closed-loop systems—NLP-driven curation, translation, and application—promise ethical, inclusive advancements by 2030.

References

[1]. Glaser, J., et al. (2021). *High-throughput virtual laboratory for computational materials discovery*. Retrieved from https://www.osti.gov/pages/biblio/1772255

[2]. Gavhane, S., et al. (2025). *Indian Knowledge Systems: Multidisciplinary research, explorations, possibilities, and opportunities*. Retrieved from https://cipuglobal.org/indian-knowledge-systems-iks-multidisciplinary-research-explorations-possibilities-and-opportunities

[3]. *NLP Notes: Processing Indian Languages*. (2025). SITAMS. Retrieved from https://sitams.ac.in/wp-content/uploads/2025/10/NLP-notes.pdf

[4]. Sain, S. K., et al. (2025). *The role of artificial intelligence in preserving and promoting Indian Knowledge Systems*. Retrieved from https://shodhsamagam.com/uploads/issues_tbl/1750496280he-Role-of-Artificial-Intelligence-in-Preserving-and-Promoting-Indian-Know

[5]. *IKS in AI*. (2025). VIIRJ Special Issues. Retrieved from https://www.viirj.org/specialissues/2025/SP2502/29.pdf

[6]. Kathavate, P. N. (2025). *Indian Knowledge Systems in machine learning applications*. World Scientific. Retrieved from https://www.worldscientific.com/doi/10.1142/S2737599425500173

[7]. IKS India Portal. (n.d.). Retrieved from https://iksindia.org

[8]. AI4Bharat Publications. (2024–2026). *Bhashini Initiative Reports*.