

CHAPTER-2

Language technology, natural language processing, and digital humanities for Bharatheeya languages: An observation

Dr. Kanthi Pranay Kumar

Lecturer of English,

kanthipranay@gmail.com

<https://doi.org/10.66483/jsp.2026.001.ch02>

ABSTRACT

In the present global scenario, one of the world's richest linguistic and cultural ecosystems can be found in Indian languages. The Bharatheeya (Indian) languages, with their hundreds of languages, scripts, dialects, and centuries-old literary traditions, provide an enormous storehouse of information, philosophy, history, and cultural expression. The promotion, processing, and preservation of these languages have become opportunities as well as challenges in the digital age of global era. Indian languages can flourish in contemporary computer and digital contexts thanks in large part to Language Technology, Natural Language Processing (NLP), and Digital Humanities. With particular reference to Bharatheeya languages. In this article, it examines the ideas, development, difficulties, tools, applications, and future prospects of language technology and digital humanities.

KEY WORDS: Language, Technology, Digital, Humanities, Bharatheeyatha

1 Introduction

Language is the foundation of human communication, culture, and knowledge transmission. Bharatheeya languages such as Sanskrit, Tamil, Telugu, Kannada, Malayalam, Hindi, Bengali, Marathi, Urdu, Assamese, Odia, Punjabi, and many others have rich literary traditions spanning thousands of years. These languages carry philosophical treatises, scientific texts, poetry, folklore, epics, inscriptions, and oral traditions.

With the rapid growth of digital technologies, there is an urgent need to bring these languages into computational frameworks. Language Technology and Natural Language Processing (NLP) provide the tools and methodologies to analyze, understand, and generate human language using computers. Digital Humanities, on the other hand, combines computational tools with humanities research to digitize, analyze, and interpret cultural and historical data.

The integration of these three domains is essential for ensuring that Bharatheeya languages are not marginalized in the global digital ecosystem dominated by English and a few other major languages.

2 Understanding Language Technology:

Language Technology: Language Technology refers to computational techniques and systems that process, analyze, store, retrieve, and generate human language in spoken or written form. It includes software, tools, standards, and resources that enable interaction between humans and machines using natural language.

<https://jnanasetupress.com/>

Bharatiya Bhasha Pariwar: A New Framework in Linguistics

<https://doi.org/10.66483/jsp.2026.001>

Components of Language Technology: Key components include: like Text message processing systems, Language Speech recognition and synthesis, Machine translation of language, Data Information retrieval, Information Text-to-Speech (TTS) and speech-to-text (STT), Optical Character Recognition (OCR), Lexical databases and corpora etc. For Bharatheeya languages, language technology must handle diverse scripts, complex morphology, free word order, and semantic richness.

3 Natural Language Processing (NLP):

Natural Language Processing (NLP) is a main branch of Artificial Intelligence (AI) and computational linguistics of language that enables the computers to understand the process, interpretation, and to generate human language in a meaningful way. NLP bridges the gap between human communication and machine understanding by applying linguistic rules, statistical models, and machine learning techniques.

Core Levels of Natural Language Processing (NLP):

- Phonology and Phonetics – Sound systems of language.
- Morphology – Structure of words of language.
- Syntax – Sentence structure of language.
- Semantics – Meaning of words and sentences of language.
- Pragmatics – Contextual meaning of language.
- Discourse – Coherence across sentences of language.
- Bharatheeya languages often exhibit rich morphology and agglutination, making NLP tasks more complex compared to English.

4 Linguistic Characteristics of Bharatheeya Languages:

Diversity of Scripts: India uses multiple scripts such as: Devanagari, Tamil, Telugu, Kannada, Malayalam, Bengali, Gurmukhi, Odia and Perso-Arabic (Urdu). Each script has unique orthographic rules, posing challenges for OCR, font standardization, and text rendering.

Morphological Richness: Most Indian languages are:

- Inflectional (Sanskrit)
- Agglutinative (Telugu, Tamil, Kannada, Malayalam etc.)

A single word can encode tense, gender, number, case, and mood, making tasks like tokenization and part-of-speech tagging difficult.

Free Word Order: Unlike English, many Bharatheeya languages allow flexible word order, complicating syntactic parsing and machine translation.

5 Digital Humanities:

Digital Humanities (DH) is an interdisciplinary field that applies digital tools, computational methods, and data-driven approaches to humanities disciplines such as literature, history, philosophy, linguistics, archaeology, and cultural studies.

Objectives of Digital Humanities: like Digitization of manuscripts and archives, Computational analysis of literary texts, Preservation of endangered languages, Cultural analytics and visualization, Public access to heritage resources. For Bharatheeya languages, DH is vital for preserving ancient manuscripts written on palm leaves, birch bark, and stone inscriptions.

6 Digitization of Bharatheeya Language Resources

Language Manuscripts and Archives: India possesses millions of manuscripts in Sanskrit and regional languages. Digitization initiatives focus on such as Scanning, Metadata creation, Transcription, Translation

Optical Character Recognition (OCR): OCR for Indian scripts faces challenges due to: like Complex ligatures, Handwritten texts, Degraded manuscripts. Nevertheless, OCR is crucial for converting printed and handwritten materials into machine-readable text.

7 Language Resources for Bharatheeya Languages:

Corpora: A corpus is a large collection of texts used for linguistic analysis and NLP training. Types include: Parallel corpora, Monolingual corpora, Spoken corpora, Annotated corpora etc.

Lexical Resources: Dictionaries, Wordnets, Thesauri, Ontologies etc.

The **Indo Word Net** is a significant resource linking lexical concepts across Indian languages.

8 Machine Translation for Indian Languages:

Machine Translation (MT) is one of the most impact NLP applications in India due to linguistic diversity. Uses grammatical and linguistic rules. Effective for closely related languages but requires extensive linguistic expertise.

Statistical and Neural MT: Modern systems use machine learning of language and deep learning of language techniques to learn translation patterns from large datasets.

Challenges: Lack of large parallel corpora, Domain-specific vocabulary, Code-mixing and code-switching etc.

9 Speech Technologies:

Speech Recognition: It Converts spoken language into text message. It is useful for: Voice assistants, Transcription, Accessibility tools etc.

Speech Synthesis: Text-to-speech systems enable digital content access for visually impaired users and language learners. Developing high-quality speech systems for Bharatheeya languages requires: like Phonetically balanced datasets, Dialect modelling

10 Applications of NLP and DH for Bharatheeya Languages

Education: such as E-learning platforms, Automated grading, Language learning apps

Governance: e-Governance Portals, Multilingual Public Services, Document Translation

Cultural Preservation: Digital archives, Folklore documentation, Oral history projects

Media and Entertainment: Subtitling, Content recommendation, Sentiment analysis of regional media

11 Challenges and Limitations

Resource Scarcity: Many Indian languages are low-resource languages with limited datasets.

Standardization Issues – like Multiple spellings, Dialectal variations, Script variations

Technical and Financial Constraints: Developing high-quality tools requires: such as Expertise, Funding, Long-term institutional support

12 Government and Institutional Initiatives

India has launched several initiatives to promote language technology: like Language technology missions, Digital libraries, AI-based translation platforms. Academic institutions, research centers, and open-source communities play a vital role in advancing these efforts.

13 Future Directions

AI and Deep Learning of Language: Transformer-based models and the large language models offer new opportunities for Indian languages.

Inclusive Digital Growth: Ensuring that rural and marginalized communities benefit from language technologies.

Interdisciplinary Collaboration: Collaboration between linguists, computer scientists, historians, and social scientists is essential.

14 Conclusion

Language Technology, Natural Language Processing, and Digital Humanities are indispensable for the survival, growth, and global presence of Bharatheeya languages in the digital era. By combining traditional linguistic knowledge with modern computational techniques, it is possible to preserve ancient wisdom while enabling innovative applications for contemporary society. The future of Bharatheeya languages depends on sustained research, inclusive policies, and collaborative efforts across disciplines. As India moves toward a digitally empowered society, empowering its languages through technology is not merely the technical necessity but the cultural responsibility.

References:

- Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing* (3rd ed.). Pearson.
- Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.
- Bharati, A., Chaitanya, V., Kulkarni, A. P., & Sangal, R. (1995). *Natural Language Processing: A Paninian Perspective*. Prentice Hall of India.
- Bharati, A., Sangal, R., Sharma, D.M., & Bai, L. (2016). *AnnCorra: Treebanks for Indian Languages*. Central Institute of Indian Languages.
- Kachru, B.B., Kachru, Y., & Sridhar, S.N. (2008). *Language in South Asia*. Cambridge University Press.
- Steever, S. B. (2019). *The Dravidian Languages* (2nd ed.). Routledge.
- Sangal, R., Malik, S., & Kumar, A. (2011). Language Technology Development in India. *Journal of the Indian Institute of Science*, 91(3), 421–440.
- McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.
- Schreibman, S., Siemens, R., & Unsworth, J. (2016). *A New Companion to Digital Humanities*. Wiley-Blackwell.
- Tharoor, S. (2018). *The Indian Origins of Digital Humanities*. Oxford University Press.
- Saravanan, K., & Natarajan, P. (2014). OCR for Printed Tamil Text Using Unicode. *International Journal of Computer Applications*, 89(2), 1–6.
- Government of India. (2020). *National education policy 2020*. Ministry of Education.
- TDIL Programme. (2019). *Technology Development for Indian Languages*. Ministry of Electronics and Information Technology, Government of India.
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *Proceedings of ACL*, 6282–6293

YOUR DISSERTATION DESERVES RECOGNITION, NOT A PLACE IN A STOREROOM

Publish Your Dissertation as a Book through the Emerging Scholars Series

Transform your thesis or dissertation into a professionally published scholarly work, boosting its academic visibility and impact.

Benefits

-  ISBN Registration*
-  Crossref DOI Assignment
-  ORCID Integration
-  Google Scholar Discoverability
-  Enhanced Citation Opportunities
-  Wider Academic Reach and Recognition
-  Long-Term Preservation and Accessibility

Publication Options

Jnanasetu Repository	Jnanasetu Repository Publication with an ISBN
✓ Crossref DOI Assignment ✓ Online Archiving and Preservation ✓ Academic Visibility ✓ Citations	✓ ISBN Registration ✓ Crossref DOI Assignment ✓ Professional Book Publication ✓ Global Accessibility ✓ Academic Visibility ✓ Citations

Looking to Publish Research Articles?

Researchers and scholars may also submit manuscripts to **Jnanasetu Journals** for quality-focused scholarly publication.

- ✓ Peer-Reviewed Journals
- ✓ DOI Assignment
- ✓ Open Access Publishing
- ✓ Academic Visibility
- ✓ Timely Editorial Processing

We Also Publish

-  Edited Books
-  Scholarly Monographs
-  Conference Proceedings
-  Academic Journals

For more details:

Press	jnanasetupress.com	support@jnanasetupress.com
Repository	jnanaseturepository.com	editor@jnanaseturepository.com
Journals	jnanasetujournals.com	editor@jnanasetujournals.com