ARTIFICIAL INTELLIGENCE AND THE PRESERVATION OF ENDANGERED INDIAN LANGUAGES

Dr. Sarita Uttamramrao Chandankar (Chapke)
Dept. of English, Smt.Vastalabai Naik Mahila Mahavidyalaya Pusad. Dist.Yeotmal
chapke.sarita5@gmail.com

Abstract
India possesses an extraordinary linguistic heritage, reflected in the fact that its people speak more than 19,500 dialects and around 780 distinct languages. These languages are not merely tools of communication but repositories of history, philosophy, ecological wisdom, folklore, and identity. Yet, linguistic diversity in India is under severe threat. UNESCO estimates that nearly 197 Indian languages are endangered, with dozens having already disappeared in the past century. Many of these endangered languages are spoken by tribal and indigenous groups, whose cultural traditions and oral knowledge systems are inextricably tied to their languages. The extinction of a language signifies not only the loss of words and grammar but also the erasure of entire worldviews. At the same time, rapid advancements in Artificial Intelligence (AI) have opened new possibilities for reversing this decline. Natural Language Processing (NLP), machine translation, speech recognition, and digital archiving tools are increasingly being used to document and revitalize languages worldwide. In India, however, most endangered languages are "low-resource," meaning they lack sufficient digital data to train robust AI models. This paper explores how AI can be harnessed to preserve endangered Indian languages while addressing challenges of data scarcity, ethical concerns, and the need for community participation. By analyzing data from linguistic surveys, digital archives, and existing AI projects, this study argues that AI can serve as a transformative tool for linguistic preservation—provided it is deployed in culturally sensitive and participatory ways.

Keywords: *Artificial Intelligence, Endangered Languages, Linguistic Diversity, Revitalization, Digital Preservation, Oral Traditions etc*

Introduction

India's linguistic diversity is both vast and fragile. The constitution officially recognizes 22 scheduled languages, but beyond these lie hundreds of others spoken by smaller communities, especially tribal and indigenous groups. These languages often lack a written script, existing primarily in oral form. For centuries, they have carried myths, legends, ritual practices, agricultural knowledge, and ecological insights. For example, the Gondi language of central India encodes detailed knowledge about forest management and biodiversity, while Khasi oral narratives preserve ancestral wisdom about community governance.

Despite their richness, many such languages are endangered. The forces of globalization, migration, and the dominance of Hindi and English in education and employment have weakened intergenerational transmission. Parents often encourage children to learn dominant languages for upward mobility, leaving ancestral tongues to decline. Scholars warn that if present trends continue, half of India's languages may vanish in the next 50 years (Devy 45). This would mark an unprecedented cultural loss, reducing not only linguistic diversity but also the diversity of thought.

Against this background, Artificial Intelligence has emerged as a powerful tool of preservation. Globally, AI has been employed in projects such as Google's Project Euphonia, which develops speech recognition for atypical and underrepresented speech patterns, and the Endangered Languages Project, which provides digital platforms for language documentation. In India, similar approaches could revolutionize efforts to document and revitalize languages. AI tools can transcribe oral stories, build digital dictionaries, create interactive learning apps, and connect younger generations with ancestral tongues.

This paper aims to investigate the potential and limitations of AI in the Indian context. It addresses three interrelated questions:

1. How can AI overcome the problem of limited digital resources for endangered Indian languages?
2. What role do communities and cultural contexts play in AI-based preservation?
3. How can AI be integrated with state policies and grassroots activism to ensure sustainable linguistic revitalization?

By answering these questions, present paper situates AI not just as a technological solution but as part of a larger cultural and ethical framework for language preservation.

Literature Review

Research on endangered languages in India emphasizes their cultural importance and vulnerability. Anvita Abbi's work on the Andaman Islands demonstrates how languages encode ecological and cultural knowledge that cannot be translated without distortion (Abbi 87). She argues

that when a language disappears, humanity loses unique insights into the relationship between people and nature. Similarly, Ganesh Devy's monumental *People's Linguistic Survey of India* documents over 780 languages, many of which are spoken by small communities at risk of extinction. Devy stresses that the decline of languages is linked to socio-economic marginalization, as communities abandon their native tongues in pursuit of survival in dominant-language environments (Devy 45).

In the field of computational linguistics, Rajesh Kumar outlines challenges specific to Indian languages, noting the lack of digitized corpora, inconsistent orthographies, and the complexity of morphologically rich languages (Kumar 214). Despite these obstacles, Kumar highlights the promise of machine learning for resource-poor languages, especially when combined with linguistic insights.

Prashant Joshi builds on this by exploring transfer learning, a technique where AI models trained on resource-rich languages are adapted to similar low-resource ones. For example, Hindi-trained models can assist Bhojpuri or Maithili, while Tamil models can support lesser-documented Dravidian languages (Joshi 56). This suggests that India's linguistic families can benefit from computational interconnections.

Other scholars highlight AI's role in revitalization. Neha Gupta, in her study of digital storytelling, shows that AI-driven interactive platforms can motivate children and diaspora communities to learn ancestral languages. By linking cultural stories with gamified language lessons, AI not only teaches vocabulary but also re-establishes cultural pride (Gupta 141).

Collectively, these works reveal a consensus that while language endangerment is a pressing issue, AI offers innovative solutions. Yet, scholars also emphasize the risks of homogenization, data exploitation, and cultural misrepresentation if AI tools are developed without community involvement.

Data Sources

The data for this paper draws from both institutional and digital resources. UNESCO's *Atlas of the World's Languages in Danger* provides a global framework for categorizing languages into vulnerable, endangered, or critically endangered. According to the Atlas, more than 197 Indian languages fall into one of these categories (UNESCO 14).

Within India, the People's Linguistic Survey of India (PLSI), led by Ganesh Devy, offers detailed accounts of speaker populations, cultural contexts, and transmission patterns. Government initiatives such as the Scheme for Protection and Preservation of Endangered Languages (SPPEL) add statistical depth, although their implementation has been uneven.

On the digital side, the Linguistic Data Consortium for Indian Languages (LDC-IL) provides corpora for computational research, though its coverage is still limited to relatively larger languages. Online platforms like the Endangered Languages Project and Google's Project Euphonia showcase global applications of AI that could be localized in India.

In addition, digital ethnographic methods, such as examining YouTube channels, Facebook groups, and WhatsApp communities where minority languages are actively used, serve as real-time data sources. These spaces demonstrate how technology already supports language survival informally.

Analysis and Discussion

Challenges of Low-Resource Languages

The primary challenge for AI in India is the scarcity of digital resources. Unlike English or Mandarin, which dominate AI research with massive datasets, tribal and minority Indian languages often lack even basic digital text collections. For instance, languages like Bhili or Kui are rarely published in digital form, making it difficult to train NLP models. As Kumar points out, computational tools require at least several thousand digitized sentences to begin functioning effectively (Kumar 215).

Transfer Learning and Multilingual Models

Transfer learning offers a partial solution. By leveraging similarities across linguistic families, AI can reduce data dependency. For example, pre-trained models in Hindi can be adapted to Bhojpuri or Magahi due to shared grammar and vocabulary. Joshi notes that such adaptations significantly reduce costs and time compared to developing entirely new models (Joshi 59). Similarly, Dravidian languages like Telugu and Kannada can support minority Dravidian tongues such as Tulu. Advances in multilingual transformer models like BERT and GPT also enhance cross-linguistic adaptability.

Preserving Oral Traditions

Many endangered Indian languages are primarily oral, which poses both a challenge and an opportunity. AI-powered speech recognition systems can transcribe oral narratives, rituals, and folk songs into digital archives. This is particularly valuable because oral traditions carry cultural knowledge that written documentation cannot fully capture. For instance, Santali folktales, once at risk of disappearing, have been preserved in digital

archives thanks to audio recording projects supported by linguistic NGOs (Abbi 89).

Automatic Speech Recognition (ASR) tools can be trained to recognize these oral languages, enabling the creation of audio dictionaries and searchable oral archives. This not only preserves language but also empowers future generations to access ancestral knowledge.

Revitalization through AI Tools

AI also supports revitalization by making endangered languages usable in modern contexts. Chatbots, translation apps, and gamified language-learning platforms allow younger speakers to practice their ancestral languages. For example, diaspora children who may never visit their ancestral villages can interact with AI-based learning tools that teach them vocabulary through cultural stories and songs. Gupta emphasizes that such tools strengthen identity and pride among younger generations (Gupta 143).

Ethical Concerns

However, ethical challenges remain. Data collection must involve informed consent from communities. Without this, AI risks repeating colonial patterns of knowledge extraction, where indigenous voices are excluded from decisions about their cultural heritage (Devy 33). Moreover, AI algorithms may homogenize linguistic variation, prioritizing dominant dialects over minority ones within the same language group. A community-driven approach is thus essential, ensuring that AI projects reflect the diversity of dialects and contexts.

Policy Integration

Finally, AI must be integrated with government initiatives. Projects like SPPEL, which aim to document endangered languages, often face funding and implementation gaps. If combined with AI-driven tools, these projects could accelerate language preservation. For instance, SPPEL's recordings could be processed using ASR to create searchable digital archives accessible to both scholars and communities.

Conclusion

The study demonstrates that Artificial Intelligence holds immense potential for the preservation and revitalization of endangered Indian languages. By addressing challenges of digitization, leveraging transfer learning, and supporting oral traditions, AI can transform the way languages are documented and used. Moreover, AI-based revitalization tools can re-engage younger generations, strengthening cultural pride and identity.

Yet, technology is not a standalone solution. The future of India's linguistic diversity depends equally on ethical practices, participatory frameworks, and strong policy support. AI must not replace communities but empower them to sustain their own languages. The collaboration of linguists, technologists, policymakers, and native speakers is crucial to ensuring that India's endangered languages continue to thrive.

In conclusion, AI should be seen as a bridge between tradition and modernity, offering hope that India's vast linguistic heritage — once feared to be on the brink of extinction — can resonate vibrantly in the digital age.

References

1. Abbi, Anvita. Endangered Languages of the Andaman Islands. Mouton de Gruyter, 2006.
2. Devy, Ganesh N. The Languages of India: Cultural and Historical Perspectives. Orient Blackswan, 2018.
3. Gupta, Neha. "Digital Storytelling and Indigenous Language Learning." Journal of Computational Linguistics and Education, vol. 14, no. 2, 2021, pp. 135–148.
4. Joshi, Prashant. "Transfer Learning for Low-Resource Indian Languages." International Journal of Artificial Intelligence and Applications, vol. 9, no. 1, 2020, pp. 55–62.
5. Kumar, Rajesh. Computational Approaches to Indian Languages. Springer, 2019.
6. UNESCO. Atlas of the World's Languages in Danger. UNESCO Publishing, 2010.