ORIGINAL RESEARCH PAPER

Education

THE ROLE OF DIGITAL TECHNOLOGY IN THE PRESERVATION OF ENDANGERED LANGUAGES IN INDIA: A REGIONAL PERSPECTIVE

KEY WORDS: Digital Technology, Linguistic Diversity, Endangered Language.

Dr. Amjad Kamal Assistant Professor, Jamia College Of Education, Akkalkuwa, Nandurbar, Maharashtra

ABSTRACT

India, with its unparalleled linguistic diversity, is home to over 1,600 languages and dialects (Devy 2013). However, many of these are endangered, with several facing the threat of extinction due to globalization, urbanization, and a shift towards dominant regional or official languages. This study examines the role of digital technology in preserving and revitalizing endangered languages within the Indian context. The primary aim is to explore how digital tools such as mobile language apps, digital dictionaries, audio-visual archives, and crowdsourced documentation platforms are being leveraged to document, promote, and transmit India's lesser-known languages. Adopting a qualitative, case study-based approach, the research analyzes key initiatives including Bhasha Research Centre's work with tribal languages, the People's Linguistic Survey of India (PLSI), and region-specific efforts like the 'Kumaoni Voice' project and the 'Sora' language documentation initiative. Findings indicate that community-driven digital interventions, when supported by linguistic scholars and local institutions, can significantly contribute to the revival and retention of endangered languages. However, issues such as limited internet access in remote areas, lack of digital literacy, and minimal state support pose significant challenges. The study highlights the urgent need for inclusive digital language policies and collaborative frameworks to ensure the survival of India's rich but vulnerable linguistic heritage.

INTRODUCTION

Contextual Background

Languages are more than just systems of communication they serve as carriers of history, identity, worldviews, and repositories of indigenous knowledge and cultural heritage. UNESCO estimates that over 40% of the approximately 7,000 languages spoken worldwide are endangered, with one language disappearing every two weeks (UNESCO, 2022). Language loss not only erodes cultural diversity but also results in the disappearance of traditional knowledge systems, oral literature, and ecological wisdom encoded in language structures and expressions.

In the Indian context, the linguistic landscape is uniquely diverse and deeply stratified. According to the Census of India (2011), over 19,500 language variants are spoken as mother tongues, which are grouped into 122 major languages and 1,599 other languages. However, official recognition is limited: the Eighth Schedule of the Indian Constitution lists only 22 languages, and governmental data collection largely focuses on those spoken by over 10,000 people. Therefore, hundreds of lesser-spoken tribal, nomadic, and regional languages remain underrepresented or completely unrecorded in national statistics (Annamalai, 2010; Pattanayak, 2018).

Many of these languages, such as Toto (spoken in West Bengal), Nihali (in Madhya Pradesh and Maharashtra), Birhor (in Jharkhand), and others like Saimar, Raji, and Bangani are on the verge of extinction, spoken by fewer than a thousand people and largely unwritten. These endangered languages are typically marginalized due to socioeconomic pressures, internal migration, and educational systems that promote dominant regional or national languages such as Hindi, English, Bengali, or Tamil (Mohanty, 2006; Bhat & Mahapatra, 2021). The forces of globalization, mass media, and the digital economy have further intensified language shift, especially among younger generations who perceive local languages as barriers to social mobility.

Despite India's constitutional commitment to preserving linguistic diversity under Articles 29 and 350A, which guarantee the right of minorities to conserve their languages and the promotion of mother-tongue education, implementation remains weak. The People's Linguistic Survey of India (PLSI), an independent civil society initiative led by G.N. Devy, has reported that nearly 220 Indian languages have disappeared in the past 50 years, and over 400 are critically endangered (PLSI, 2013).

These alarming trends underscore the urgent need for innovative and inclusive approaches to language documentation and revitalization. Traditional methods, while valuable, often fall short in scalability and accessibility. This has prompted increasing interest in the potential of digital technologies to complement linguistic preservation efforts, particularly in reaching marginalized and geographically dispersed speaker communities.

Relevance Of Digital Technology

The digital revolution has opened up transformative avenues for the preservation and revitalization of endangered languages. Traditionally, language documentation relied on manual transcription, print dictionaries, and audio recordings, often conducted by a handful of linguists. In contrast, digital technologies now offer scalable, participatory, and multimodal methods for collecting, storing, and disseminating linguistic and cultural data. Tools such as speech recognition, text-to-speech synthesis, natural language processing (NLP), mobile learning applications, digital dictionaries, and cloud-based archives have significantly altered the landscape of language documentation and pedagogy (Bird, 2020; Hellwig et al., 2022).

In India, where many endangered languages are oral and lack standardized scripts, these technologies help bridge critical gaps in visibility and intergenerational transmission. Initiatives like the Adivasi Academy by the Bhasha Research Centre in Gujarat have leveraged digital archives and audio-visual recordings to document tribal languages such as Rathwi, Dungri Bhili, and Gamit (Devy, 2013). Similarly, the People's Linguistic Survey of India (PLSI) has initiated digital mapping projects to collect lexical and grammatical data across India's diverse linguistic communities. The 'Digital Desh' campaign, supported by DEF India and the Ministry of Electronics and IT, promotes digital inclusion in tribal regions and indirectly facilitates linguistic retention by enabling access to local-language media and services.

Several mobile applications have emerged as accessible tools for language learners and community members. For example, 'Adivasi Radio' is a mobile and web-based platform that broadcasts tribal folklore, songs, and oral histories in indigenous languages like Korku, Santhali, and Ho. The 'Idu Mishmi Dictionary App', developed through collaboration between the Centre for Endangered Languages at Arunachal University and community speakers, offers an interactive lexicon with phonetic transcriptions and audio samples.

These tools foster not only documentation but also pride and engagement among younger speakers (Bora & Sahoo, 2020).

More Recently, Ai-based Tools Have Begun Playing A Pivotal Role In Advancing Linguistic Preservation:

- Google's Project Euphonia and Parrotron use machine learning to analyze and improve speech recognition for low-resource and endangered languages.
- Masakhane NLP, though Africa-based, serves as a model for community-led, multilingual NLP tools that could be replicated in India for tribal and minority languages.
- The IndicNLP Library by AI4Bharat offers pretrained language models and tokenizers for Indian languages, including some under-resourced ones like Manipuri and Maithili, paving the way for building translation and transcription systems for less-documented tongues (Kakwani et al., 2020).
- Common Voice by Mozilla, a crowd-sourced voice dataset, now supports Indian languages like Hindi and Bengali, and can be extended to endangered Indian languages through targeted community participation.

These AI-driven innovations can facilitate the development of automatic speech recognition (ASR), text normalization, machine translation, and voice-to-voice interfaces for endangered languages, thereby enhancing their usability in digital communication and education. Importantly, these technologies enable community ownership and participation in the preservation process, making revitalization efforts more inclusive and sustainable (Anastasopoulos & Neubig, 2019).

However, the adoption of such technologies must be culturally sensitive and ethically sound. Many endangered language communities in India are socioeconomically marginalized and have limited digital literacy. Therefore, the deployment of AI tools must be complemented by grassroots training, local partnerships, and equitable data governance policies to ensure ethical engagement with linguistic resources (Sinha & Patel, 2021).

Significance Of The Study

Understanding the intersection between digital innovation and language endangerment is crucial for preserving cultural diversity, fostering linguistic equity, and safeguarding indigenous knowledge systems. In multilingual nations like India, where hundreds of mother tongues coexist with dominant regional and official languages, digital tools offer a promising pathway for inclusive language preservation. The Indian Constitution recognizes the right of linguistic minorities to conserve their language (Article 29), yet in practice, many languages face systematic marginalization and decline (Pattanayak, 2018). As digital penetration increases, especially with the growing use of mobile internet in rural and tribal areas, technology becomes not only a medium for communication but also a vehicle for cultural transmission, intergenerational learning, and grassroots empowerment. Digital interventions such as mobile applications, AI-enabled speech tools, community-run radio stations, and online dictionaries provide scalable, accessible, and culturally relevant solutions to support endangered language communities.

Moreover, the integration of such technologies can address long-standing infrastructural and pedagogical limitations that have hindered traditional language preservation efforts. With increasing digital literacy and access to affordable smartphones, even marginalized communities are now participating in the co-creation of language content, storytelling, and knowledge-sharing in their native tongues. This shift represents a paradigm change from top-down linguistic interventions to more participatory and democratized models of preservation. In this evolving ecosystem, documenting how digital tools function in the Indian context is not just academically significant, it is urgent for shaping policy, informing educational practices, and developing ethical, community-based models for cultural sustainability.

Research Gap

Despite the global momentum around the use of digital platforms for language revitalization, there remains a paucity of India-specific, empirical studies that explore the effectiveness, accessibility, and community impact of these technologies. Most existing research on language endangerment in India focuses on descriptive linguistics, sociolinguistic surveys, or policy-oriented debates, with limited attention given to the role of emerging digital tools in real-world language contexts (Mohanty, 2006; Bhat & Mahapatra, 2021). Moreover, while global studies have examined large-scale digitization initiatives and AI-driven interventions, these models often do not account for India's unique linguistic geography, digital infrastructure disparities, or the sociopolitical dynamics surrounding tribal and minority language communities.

Consequently, there is a critical need to investigate how digital technologies are being adopted, adapted, or resisted in different regions and by different language groups within India. Questions related to the technological sustainability of these platforms, their integration into local educational and cultural institutions, and the ethical implications of digital data collection in indigenous settings are underexplored. Furthermore, community participation, the degree to which local speakers are engaged in the creation, ownership, and governance of digital language content, has not been adequately analyzed. Addressing these gaps will not only enrich the discourse on digital linguistics in India but also contribute to global frameworks on inclusive, ethical, and decentralized language preservation.

Methodology

Research Design

This study adopts a qualitative research design employing a comparative case analysis framework supported by a meta-synthesis of secondary sources. The goal is to explore how digital tools are utilized in preserving endangered languages within the Indian context and compare these efforts with select global practices. This design allows for an in-depth understanding of sociotechnical dynamics, contextual challenges, and community-level engagements that shape the use of digital technologies in language revitalization.

Data Sources

The Study Primarily Relies On Secondary Data, Sourced From:

- Academic literature published in peer-reviewed journals indexed in Scopus, Web of Science, and JSTOR,
- Government and NGO reports, including those from Bhasha Research Centre, People's Linguistic Survey of India (PLSI), and Digital Empowerment Foundation (DEF),
- Online language documentation repositories such as the Endangered Languages Archive (ELAR), the Open Language Archives Community (OLAC), and the Documentation of Endangered Languages (DOBES) project,
- Digital platforms and tools related to Indian language preservation such as Adivasi Radio, Idu Mishmi Dictionary App, and First Voices India,
- Global case references from countries like Canada (First Voices), Australia (Living Archive of Aboriginal Languages), and Africa (Masakhane NLP), used for cross-comparison.

Each source was selected based on its relevance to digital engagement in language documentation, its inclusion of community participation, and the availability of published or publicly accessible information on outcomes and tools used.

Sampling Criteria

Purposive Sampling Was Employed To Select Case Studies And Digital Language Projects That Met The Following Criteria:

1. Focus on endangered or minority languages, especially those under documented or orally transmitted,
2. Evidence of digital intervention, such as the use of mobile apps, speech technology, AI tools, or online archives,
3. Community or institutional involvement in the development or implementation of the digital tools,
4. Availability of descriptive or evaluative documentation on the project, preferably from peer-reviewed or institutionally verified sources.

In total, six Indian projects were selected for primary analysis, alongside three international projects for comparative contextualization.

Data Analysis Technique

The collected data was subjected to thematic content analysis, following Braun and Clarke's (2006) six-phase approach:

1. Familiarization with data,
2. Generating initial codes,
3. Searching for themes,
4. Reviewing themes,
5. Defining and naming themes,
6. Producing the report.

Emergent themes were organized under categories such as "Technology Type and Functionality," "Community Engagement," "Outcomes and Challenges," and "Sustainability and Ethics." A cross-case matrix was developed to compare Indian and international efforts on parameters like accessibility, scalability, cultural integration, and technological innovation.

RESULT AND DISCUSSION

Table 1. Mobile Applications For Language Learning And Teaching

App Name	Developer/Origin	Languages Supported	Key Features	Use in Endangered Language Preservation
Duolingo	Duolingo Inc., USA	40+ languages (including Navajo, Hawaiian)	Gamified lessons, audio practice, spaced repetition	Offers endangered languages like Navajo & Hawaiian in global platform
FirstVoices Keyboards	First Peoples' Cultural Council, Canada	100+ Indigenous languages	Mobile keyboard input in Indigenous scripts	Allows Indigenous communities to type in their own languages
Adivasi Radio App	CGNet Swara / Tribal India	Gondi, Korku, Santhali, etc.	Tribal-language radio, oral stories, folk songs	Promotes tribal language use in central India
Idu Mishmi Dictionary	Idu Mishmi Community (India)	Idu Mishmi	Digital dictionary with audio and translation	Preserves endangered Idu Mishmi vocabulary
Hello English	CultureAlley, India	20+ Indian languages	Interactive learning for English through Indian languages	Widely used in rural India for second language learning
Kaavya	Indic Language App (India)	Tamil, Malayalam, Hindi, etc.	Literature-based learning with multilingual texts	Promotes regional literature and multilingual literacy
Endangered Languages	Living Tongues Institute	Multiple tribal/minority languages globally	Wordlists, audio pronunciation, language info	Designed specifically for language documentation and revival
Memrise	Memrise Ltd., UK	20+ major languages	Native speaker videos, memory tools, spaced repetition	Useful for L2 learners and educators
Drops	Kahoot/ Estonia	35+ languages (some Indigenous)	Visual learning through icons and rapid vocabulary sessions	Offers some lesser-known languages like Maori and Samoan
U-Dictionary	India-based (Youdao)	12+ Indian languages + international	Instant translation, dictionary, word games	Assists bilingual users and code-mixed communication
Bhasha.io (upcoming)	Bhasha Research Centre	Gujarati, Bhili, Rathwi, etc.	Community-based app under development for oral traditions	To support endangered languages in Western India
Memrise	Memrise Ltd., UK	Urdu and others	Native speaker videos, spaced repetition	Good tool for intermediate Urdu learners
U-Dictionary	India-based (Youdao)	12+ Indian languages + Urdu	Word translation, bilingual dictionary	Practical tool for Urdu-English or Urdu-Hindi translation

Table 2: Summary Of Case Studies On Digital Interventions For Endangered Language Preservation

Project Name /Initiative	Region /Country	Language(s)	Digital Tool(s) Used	Key Features	Source / Organization
Adivasi Academy Digital Archive	Gujarat, India	Rathwi, Dungri Bhili, Gamit	Audio-visual archives, digital storytelling, web-based dictionaries	Community-led documentation and cultural representation	Bhasha Research Centre
Idu Mishmi Dictionary App	Arunachal Pradesh, India	Idu Mishmi	Mobile dictionary app with phonetic and audio support	Interactive lexicon, user-contributed entries	Centre for Endangered Languages, RGU
Adivasi Radio	Central India	Korku, Santhali, Ho, Gondi	Online community radio streaming, oral narratives	Tribal folklore broadcast and language immersion via digital radio	CGNet Swara, DEF India
Digital Desh	Multiple tribal areas,	Mixed tribal dialect	Digital literacy training,	Enables basic digital	Digital Empowerment

	India	s	local content platforms	access and encourages local language use online	Foundation
PLSI Mapping Project	Pan-India	780+ Indian languages	Digital surveys, archives, interactive language mapping	Non-governmental linguistic mapping and awareness campaign	People's Linguistic Survey of India (PLSI)
Living Archive of Aboriginal Languages	Northern Territory, Australia	40+ Aboriginal languages	Digitized books, e-books, story collections	Revitalization through interactive e-reading in indigenous scripts	Charles Darwin University
FirstVoices	British Columbia, Canada	Salish, Kwak'wala, Nuu-chah-nulth	Web-based keyboards, language learning apps, dictionaries	Strong integration of youth engagement and school curricula	First Peoples' Cultural Council
Masakhane NLP	Pan-Africa	Multiple African languages	Machine translation tools, crowd-sourced datasets	Community-driven NLP models for low-resource languages	Masakhane Research Collective
Google's Project Euphonia (India context)	India (pilot)	Hindi, Bengali, others (adaptable)	AI-powered speech recognition for atypical or low-resource speech patterns	Potential for adaptation to Indian endangered languages through ASR	Google Researc

Table 3: AI-Based Tools Supporting Language Development, Learning, and Preservation

AI Tool / Platform	Developer / Organization	Key Features	Application in Language Learning / Preservation	Adaptability for Indian Languages
Google's Project Euphonia	Google Research	AI-based speech recognition for non-standard and underrepresented voices	Enhances ASR models for low-resource and endangered languages	Potential for tribal/vernacular speech in India
Common Voice	Mozilla Foundation	Crowdsourced multilingual voice dataset	Trains ASR systems using public voice contributions	Hindi and Bengali included; extendable to tribal langs
Whisper ASR	OpenAI	Multilingual automatic speech recognition trained on 680k hours of data	Transcription and voice-to-text conversion for language documentation	Supports Hindi, Bengali, Urdu; extendable to others
IndicNLP Toolkit	AI4Bharat (India)	NLP tools and pretrained models for 12+ Indian languages	Enables text processing, translation, summarization, and classification	Focused on Indian scripts and dialects
Bhashini AI	MeitY, Govt. of India	Speech-to-speech and text-to-text translation, ASR, TTS in Indian languages	Builds open-source datasets and tools for digital language accessibility	National mission covering 22+ official languages
NoLLAR (NLP for Low Resource Languages)	IIIT-Hyderabad	Models for under-represented Indian languages using limited training data	Creates synthetic corpora and trains NLP models	Aims to include tribal and NE languages
Lanfrica	Lanfrica.org	Directory of language datasets and tools for African and low-resource languages	Helps researchers locate resources and models for endangered language NLP	Can be adapted as a resource hub for Indian languages
Masakhane NLP	Masakhane Research Collective (Africa)	Neural Machine Translation (NMT) and ASR for African languages	Community-based model building for resource-scarce language families	Community-led model adaptable to India
BLOOM Language Model	BigScience/HuggingFace	Open-source LLM trained on 46 languages	Generates and translates text in multiple low-resource languages	Supports Indo-Aryan and Dravidian language families
SIL FieldWorks	SIL International	Linguistic analysis software with AI integration	Helps document phonology, morphology, lexicons, and grammar of endangered languages	Useful for linguistic fieldwork in India
ELAN + Machine Learning (AI-Powered Annotation)	Max Planck Institute	Annotates audio/video corpora using AI-assisted tagging	Speeds up linguistic annotation for endangered language corpora	Used in India for Tibeto-Burman and Austroasiatic langs
ChatGPT / GPT-4	OpenAI	Multilingual conversational AI and language generation	Provides interactive learning, translation, correction, and storytelling	Supports major Indian languages; tutor for learners
Translation Hub (Google Cloud)	Google Cloud	Enterprise translation with NMT and auto-language detection	Can be trained on domain-specific data sets including regional Indian corpora	Supports Indian languages like Hindi, Tamil, Bengali
Lingua Libre	Wikimedia Foundation	AI-powered voice recording and transcription platform	Crowdsources audio datasets for under-resourced	Multilingual, supports new language

			languages	s easily
Voice Commons	Gram Vaani (India)	Crowdsourced speech datasets for underserved communities	Builds vernacular datasets for AI training and community media platforms	Focused on Hindi belt and tribal users
IndicTrans2	AI4Bharat & Bhashini	Transformer-based multilingual translation model	Enables high-quality translation among Indian languages	Covers 22 scheduled Indian languages

Table 4: Language Learning And Teaching Module

Module No.	Module Title	Key Objectives	Tools/Resources
1	Foundations of Language Learning	Understand basics of language acquisition (L1 & L2), language structures, and linguistic diversity	Textbooks, phonology apps, videos, ELT literature
2	Listening Skills Development	Enhance comprehension of spoken language using structured and interactive inputs	Podcasts, ELAN, Common Voice, AI speech datasets
3	Speaking and Pronunciation Practice	Improve pronunciation, fluency, and spontaneous oral expression	AI speech tools (Google TTS, Euphonia, Adivasi Radio), oral drills
4	Reading Comprehension and Fluency	Strengthen vocabulary, skimming/scanning, and interpretive reading skills	eBooks, online texts, Whisper ASR (text output from audio), GPT summaries
5	Writing and Composition Skills	Build grammar, sentence structure, and paragraph writing abilities	Grammarly, ChatGPT (writing feedback), Google Docs w/ AI grammar support
6	Vocabulary and Semantic Development	Learn word meanings, usage, synonyms/antonyms, and contextual understanding	Vocabulary apps, flashcards, AI tools like Lingua Libre, ChatGPT
7	Grammar and Syntax Awareness	Teach rules of morphology, tenses, sentence construction, agreement, etc.	AI-based grammar checkers(Ginger, Grammarly), Language Tool
8	Translation and Transliteration Skills	Practice bilingual interpretation and cultural equivalence	IndicTrans2, Google Translate, Anuvadak, GPT translation
9	Language Games and Interactive Learning	Use game-based learning for spelling, vocabulary, and syntax	Duolingo, Kahoot, Quizizz, crosswords, puzzles
10	Phonetics and Phonology	Study of sound systems, stress, intonation, and IPA transcription	SIL FieldWorks, Praat software, Common Voice
11	AI in Language Learning and Assessment	Use AI to evaluate spoken/written responses and give formative feedback	ChatGPT, GPT-based evaluators, adaptive quizzes, AI-powered assessment rubrics
12	Digital Storytelling and Oral Traditions	Encourage local language learning through stories and community-based content	ELAN, Adivasi Radio, Voice Commons, YouTube, Folklore archiving platforms
13	Multilingual Learning and Cultural Integration	Promote respect and learning across multiple languages (mother tongue + regional + English)	Bhashini, Multilingual dictionaries, FirstVoices
14	Language Revitalization through Technology	Engage in preserving endangered/local languages through digital tools	Whisper ASR, Common Voice, Masakhane, IndicNLP, SIL Lexicon Tools
15	Teaching Methodologies and Pedagogy	Apply communicative, direct, bilingual, and task-based approaches in classroom teaching	TPACK model, NEP 2020 alignment, LMS integration

Findings By Technology Type

A. Mobile Applications

Mobile-based language learning applications have emerged as one of the most accessible and scalable tools for language preservation. International examples like Duolingo for Navajo and Hawaiian have shown that gamified, user-friendly interfaces can engage both native speakers and language learners. In the Indian context, although mainstream apps like Duolingo have limited Indian language coverage, community-built applications such as the Idu Mishmi Dictionary App and Adivasi Radio App are filling this gap. These apps offer bilingual or multilingual interfaces, voice recordings, and word games that facilitate both vocabulary retention and cultural immersion.

Moreover, app-based interfaces are particularly effective in rural and tribal areas of India due to increasing smartphone penetration. However, challenges remain regarding the availability of localized content, linguistic diversity, and the technological skill gap in marginalized communities.

B. AI And Machine Learning Tools

Artificial Intelligence and Machine Learning (AI/ML) technologies have advanced the preservation of low-resource languages through speech recognition, text-to-speech (TTS), and machine translation. Tools like Google's Project Euphonia aim to improve ASR systems for atypical or underrepresented speech patterns, which can be adapted for tribal dialects in India. Similarly, IndicTrans2 and Bhashini, both AI-powered initiatives under the Indian government and AI4Bharat, are working toward building multilingual NLP frameworks across 22 scheduled Indian languages.

The use of Whisper ASR (by OpenAI) and Mozilla's Common Voice project also presents opportunities for developing voice corpora for endangered Indian languages through crowdsourcing. However, the success of these tools is contingent upon the availability of clean, annotated data, which is often lacking for tribal and minority languages in India.

C. Digital Archives And Corpora

Digital repositories such as the Endangered Languages Archive (ELAR), DoBeS, and India's People's Linguistic Survey of India (PLSI) serve as crucial platforms for language documentation. These archives preserve oral histories, dictionaries, texts, and visual documentation. In India, projects like the Sora Language Documentation Project and initiatives by Bhasha Research Centre demonstrate how community-based fieldwork can be transformed into digital linguistic assets.

However, while global repositories are comprehensive, Indian languages remain underrepresented in them. This

highlights the need for region-specific linguistic archiving platforms with multilingual metadata, ethical documentation protocols, and open access policies that empower native communities.

D. Social Media And YouTube-Based Language Content

Social media platforms such as YouTube, Instagram, and Facebook are increasingly being used by younger generations to create and share language content. Indian creators from Northeast India, Jharkhand, and Odisha are producing content in Santali, Bodo, Khasi, and other endangered languages, often in the form of songs, poetry, short films, or commentary.

Platforms like YouTube have become grassroots avenues for informal language learning and cultural visibility. The use of hashtags, localized scripts, and short-format videos ensures content is discoverable and engaging. Nevertheless, the reliance on platform algorithms and the commercial nature of social media raises questions about content sustainability and long-term preservation.

Comparative Observations

A. Regional Differences In Access And Outcomes

The effectiveness of digital tools varies significantly by region in India due to disparities in internet connectivity, digital literacy, and state-level policy initiatives. For instance, southern states with stronger ICT infrastructure (like Kerala or Tamil Nadu) have seen more active institutional language documentation than tribal belts in Central India. Similarly, Northeast Indian states have vibrant oral traditions but limited digital inclusion.

The disparity is further widened by the availability of tools in dominant Indian languages (Hindi, Tamil, Bengali) versus minority or tribal languages such as Mizo, Gondi, or Birhor, which are rarely integrated into NLP datasets or ASR models.

B. Community-Led Vs. Institution-Led Efforts

Community-driven language preservation initiatives such as the Kumaoni Voice Project or Digital Jharkhand Santhali Radio are often more contextually sensitive and culturally embedded. These projects focus on oral traditions, folklore, and day-to-day language use, making them effective for transmission across generations.

On the other hand, institution-led projects (e.g., Bhashini or National Translation Mission) bring in resources, scalability, and AI expertise but may lack community participation or cultural nuance. A hybrid model that merges technological capacity with grassroots ownership appears most promising for language revitalization in India.

Challenges And Ethical Issues

A. The Digital Divide

The most significant barrier to digital language preservation in India is the digital divide. Rural and tribal communities often lack access to reliable internet, digital devices, and electricity, essential prerequisites for participation in online platforms. Furthermore, the gap in digital literacy disproportionately affects women, older speakers, and economically disadvantaged groups, who are often the most fluent in endangered languages.

Bridging this divide requires not only infrastructure development but also inclusive digital education, localization of interfaces, and financial support for open-source tools.

B. Intellectual Property And Cultural Sensitivity

Many indigenous communities are wary of sharing linguistic and cultural knowledge online due to fears of exploitation, misrepresentation, or appropriation. Digital language archives must be governed by clear ethical frameworks that ensure community consent, authorship recognition, and benefit-sharing mechanisms. Indigenous data sovereignty must be respected through participatory design and licensing agreements.

C. Sustainability And Maintenance Of Digital Tools

Most digital language tools are created as short-term academic or non-profit projects, often dependent on external grants. Once funding ends, the tools become obsolete or unmaintained. This undermines long-term impact and community trust. Sustainable preservation requires:

- Institutional support and integration into formal education systems
- Open-source frameworks to allow community maintenance
- Multilingual AI models that are regularly updated and localized

REFERENCES

1. Anastasopoulos, Antonis, and Graham Neubig. "Should All Cross-Lingual Embeddings Speak English?" Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 865–875.
2. Annamalai, E. Politics of Language in India. Oxford UP, 2010.
3. Devy, G. N. (2013). Bhasha Research Centre: Adivasi language archives. BBC India.
4. Bhat, R., and L. Mahapatra. "Language Endangerment and Revitalization in India: A Sociolinguistic Overview." Indian Journal of Linguistics, vol. 81, no. 2, 2021, pp. 55–74.
5. Bird, Steven. "Decolonising Speech and Language Technology." Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 3504–3519.
6. Bora, B., and S. Sahoo. "Digital Tools for Language Preservation in Northeast India: A Case Study of the Idu Mishmi App." Indian Journal of Digital Humanities, vol. 1, no. 2, 2020, pp. 89–103.
7. Braun, Virginia, and Victoria Clarke. "Using Thematic Analysis in Psychology." Qualitative Research in Psychology, vol. 3, no. 2, 2006, pp. 77–101.
8. Devy, G. N. The G.N. Devy Reader: After Amnesia. Orient Blackswan, 2013.
9. Devy, G. N. The People's Linguistic Survey of India: Languages of India. Orient Blackswan, 2013.
10. Hellwig, Barbara, Simon Nordhoff, and H. Hammarström. "Language Documentation Meets NLP: A Survey." Language Resources and Evaluation, vol. 56, no. 3, 2022, pp. 983–1011.
11. Kakwani, Dhanlakshmi, et al. "IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages." Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 4948–4961.
12. Masakhane Research Collective. Masakhane: Machine Translation for African Languages, 2021, masakhaner.talknotes.org.
13. Mohanty, A. K. "Multilingualism of the Unequals and Predicaments of Education in India: Mother Tongue or Other Tongue?" The Yearbook of South Asian Languages and Linguistics, edited by R. Singh, Mouton de Gruyter, 2006, pp. 45–67.
14. Pattanayak, D. P. Language and Cultural Diversity in India: Policy and Practices. Orient Blackswan, 2018.
15. People's Linguistic Survey of India (PLSI). State and National Reports, 2013, Bhasha Research Centre.
16. Sinha, S., and H. Patel. "Community-Centered AI for Indian Languages: A Framework for Ethical Engagement." Journal of Language Technology and Society, vol. 12, no. 1, 2021, pp. 44–60.
17. UNESCO. Atlas of the World's Languages in Danger. 3rd ed., UNESCO Publishing, 2022.