



The Role of Artificial Intelligence in the Preservation and Revitalization of Endangered Languages, with Particular Reference to Indian Languages

Iqra Parveen

PhD Scholar,
Department of Linguistics,
Aligarh Muslim University, Aligarh, India

Mohd. Yasir

PhD Scholar,
Department of Linguistics,
Aligarh Muslim University, Aligarh, India

Prof. Shabana Hameed

Professor,
Department of Linguistics,
Aligarh Muslim University, Aligarh, India

Abstract: This paper examines the transformative role of Artificial Intelligence (AI) in the preservation and revitalization of endangered languages, with particular emphasis on the Indian context. As linguistic diversity continues to decline due to globalization, urbanization, and intergenerational language shift, AI offers innovative solutions for documenting, preserving, and promoting low-resource and endangered languages. The study explores the application of AI technologies, including natural language processing (NLP), machine learning, automatic speech recognition, and neural machine translation, in language documentation, transcription, translation, corpus development, and digital language learning.

The paper analyzes how AI enhances the efficiency and scalability of language preservation compared to conventional documentation methods by enabling automated data processing, multilingual translation, and interactive educational platforms. It further examines selected case studies, including the community-



led revitalization of tereo Māori in New Zealand, AI-assisted documentation of Indigenous languages in North America, and emerging initiatives such as Bhashini and BhashaDaan in India. These examples demonstrate that AI can strengthen language revitalization when combined with community participation and culturally responsive implementation.

Recognizing that technological innovation alone cannot reverse language decline, the study also evaluates key ethical and governance issues, including data ownership, privacy, algorithmic bias, transparency, and community control over linguistic resources. It argues that sustainable language revitalization requires AI systems that complement rather than replace traditional community-led efforts. Overall, the paper highlights the potential of AI to support linguistic diversity by improving documentation, expanding educational access, and strengthening language transmission. It concludes that responsible, inclusive, and community-centred AI offers a promising pathway for safeguarding endangered languages and preserving cultural heritage for future generations.

Keywords: Artificial Intelligence; Endangered Languages; Language Revitalization; Natural Language Processing; Linguistic Diversity; Indian Languages.

Introduction

Language is the primary means by which human communities encode knowledge, carry cultural memory, and build identity across generations. Almost 7000 languages are currently in use worldwide, and a startling number of these are in danger of dying out, with predictions that a language dies every few weeks, on average, and that nearly 40 percent of the world's languages are at acute risk of extinction (UNESCO Ad Hoc Expert Group on Endangered Languages, 2003; Visual Capitalist, 2024). India is a microcosm of this crisis as it also has one of the most linguistically diverse populations in the world. In addition to the 22 languages officially mentioned in the Eighth Schedule of the Constitution, hundreds of tribal, minority, and regional languages continue to thrive with declining populations of speakers, few supporting institutions, and virtually no digital availability (Government of India, Ministry of Electronics and



Information Technology, 2025). Until now, the documentation of such languages has relied on the time-consuming work of trained linguists transcribing recordings by hand, compiling dictionaries, and writing grammars over a number of years or decades. Although this methodology is rigorous, it failed to keep pace with the rate of language attrition. However, the advent of Artificial Intelligence (AI), and specifically Natural Language Processing (NLP), Machine Learning (ML), and Automatic Speech Recognition (ASR) has started to change the game, providing tools that are able to translate, model, and transcribe languages at a scale and speed above what was once imaginable (Pinhanez et al., 2024).

Context and Significance

The importance of this inquiry lies in two present-day global issues: the decline of linguistic diversity and the rapid, sometimes uncontrolled, spread of AI technologies across cultural and social life. Language loss is not only a linguistic process but also a loss of plants and medicines taxonomy, oral history, kin relations, and worldview, which cannot be translated into the dominant languages (Bromham et al., 2022). Meanwhile, the proliferation of AI language technology is growing rapidly, from India's government-funded translation initiative Bhashini (Digital India BHASHINI Division, 2026) to community-based speech recognition efforts such as Te Hiku Media's work on tereo Māori (NVIDIA, 2024). It is important to linguists, policy makers, technologists, and communities whose languages are involved to understand how these technologies can be used responsibly, and where they are not.

AI and NLP for Language Documentation

A significant amount of recent scholarship has emerged on the use of NLP techniques to document endangered and low-resource languages. There is a large body of recent scholarship on adapting NLP techniques to document endangered and low-resource languages. Pinhanez et al. (2024) discuss a multi-year work program involving indigenous communities in Brazil in which the contributions of automated language identification, fine-tuned machine translation, and emergent “Indigenous Language Models” support indigenous written communication and documentation, explicitly foregrounding the ethical complexity of working with communities whose languages have limited digital footprints. They show that models can be adapted to low-



resource varieties with limited target-language data, as illustrated by other NLP approaches for low-resource scenarios surveyed by Hedderich et al. (2021), including cross-lingual transfer learning, data augmentation, and active learning. In addition to this technical literature, Mager et al. (2023) discuss specific challenges in applying computational linguistics to Indigenous language revitalization, suggesting that technical approaches alone do not address the need for strong community collaboration.

Students will learn about speech technologies for endangered languages

Automatic Speech Recognition (ASR) has been one of the most impactful AI tools in this field, as many endangered languages are spoken. Perhaps the most oft-quoted case is Te Hiku Media's ASR system, built to transcribe tereō Māori, which used the open-source NVIDIA NeMo toolkit and explicitly designed data-governance protocols to ensure control over voice data is always maintained by the Māori community, with a transcription accuracy of around 92 percent (NVIDIA, 2024; International Telecommunication Union, 2022). This project is often referred to as a “model of AI sovereignty” because it demonstrates that technical acumen and community control need not be mutually exclusive. Likewise, in Indian contexts, Adiga et al. (2021) introduce an ASR system and speech corpus for the classical language Sanskrit, which has reduced usage and is considered an endangered language, and AI4Bharat's (2024) IndicVoices project builds a large multilingual speech corpus of the major Indian languages for use in downstream applications such as ASR and translation.

AI in Language Learning and Revitalization Pedagogy

A second literature stream focuses on the use of AI in language teaching and transmission across generations. Commercial providers like Duolingo have integrated AI-based chatbots and adaptive learning algorithms into their endangered-language courses, such as Hawaiian and Navajo, extending access to learners beyond classroom walls (Amazon Web Services, n.d.). Meanwhile, Toppo (2025) notes that Duolingo's new “AI-first” approach—replacing its human course contributors with AI-generating technology—has raised concerns among language communities and linguists about the accuracy and cultural sensitivity of the AI-generated



content. The question of scale and quality recurs throughout the literature on AI for language education.

The Indian Context

In India, Bhashini, launched in 2022 under the National Language Translation Mission, is the most significant initiative in the field of AI language infrastructure and offers translation, speech-to-text, and text-to-speech services in the twenty-two Scheduled Languages and increasing tribal languages (Digital India BHASHINI Division, 2026; Government of India, Ministry of Electronics and Information Technology, 2025). The acute scarcity of data in the development of AI for the smaller languages of India is being tackled through complementary initiatives like BhashaDaan, which is a crowd-sourced voice-donation program by the government and stakeholders, and Scheme for Protection and Preservation of Endangered Languages (SPPEL), which was launched in the Government of India (MoEIT, 2025; Reboot Democracy, 2026). The use of AI for endangered languages in India is still not widely explored in academia, with a few studies highlighting the potential of AI tools for endangered languages in North-East Indian contexts (e.g., Haokip 2022; Kavitha et al., 2023), and urging for the deeper embedding of AI research into community-based language planning.

Ethical and Critical Perspectives

A developing body of critical literature warns of the uncritical enthusiasm for language revitalization driven by AI. Bird (2020) calls for the “decolonization” of speech and language technology, saying that in the past, NLP research had focused on Indigenous languages as objects rather than partners in technological development. Cooper et al. (2024) interviewed seventeen language-technology researchers who have worked with Aboriginal and Torres Strait Islander communities and similarly find that only when the process of engagement – as well as the artifact itself – is considered does a project actually serve a community. Indigenous data-sovereignty frameworks, such as the CARE principles for Indigenous data governance (Carroll et al., 2020) and the First Nations OCAP principles, have become key questions to consider as to whether or not an AI project respects community authority over data, particularly linguistic data (Roberts & Montoya, 2023; Cultural Survival, 2025)



Theoretical Framework

The Graded Intergenerational Disruption Scale and Fishman's Reversing Language Shift. However, among the most influential theoretical frameworks in language-revitalization scholarship is Joshua Fishman's (1991) Reversing Language Shift, accompanied by his Graded Intergenerational Disruption Scale (GIDS). Fishman's model presents the concept of language vitality along a scale of intergenerational transmission, from stage 8 (the language is remembered only by isolated elderly users) to stage 1 (the language is used in higher education, government, and mass media). That is, Fishman notes, interventions at “higher” levels (media broadcasting or government employment) are of little value if “lower” levels (at home, in the community) aren't addressed. In this paper, GIDS is used as a lens to assess the point of intervention for specific AI interventions in the transmission chain and to assess the match between the interventions and the vitality of the target language.

UNESCO's Nine Factors of Language Vitality

A complementary framework is provided by UNESCO's nine factors of language vitality and endangerment, which include intergenerational transmission, absolute number of speakers, proportion of speakers within the total population, trends in existing language domains, response to new domains and media, availability of materials for language education and literacy, governmental and institutional attitudes, community members' attitudes toward their own language, and the amount and quality of documentation (UNESCO Ad Hoc Expert Group on Endangered Languages, 2003). This study has used the factors of “response to new domains and media” and “amount and quality of documentation” most specifically, as they have been directly implicated in AI-driven interventions.

Diffusion of Innovations

The study is based on Rogers' (2003) Diffusion of Innovations theory that outlines five factors that influence the adoption of an innovation: relative advantage, compatibility, complexity, trialability, and observability. In the context of language technology, this means that AI tools are more likely to be embraced on a long-term basis in the community when they integrate well with the local culture and pedagogical practices, when they offer benefits that are



perceived as greater than those of non-digital solutions, when they are taken up by channels that are visible and trusted within the community, rather than imposed from the outside.

The AI-Mediated Revitalization Cycle is proposed as a conceptual model.

This paper proposes a conceptual model: the AI-Mediated Revitalization Cycle, consisting of five interlinked stages: (1) community-governed data collection, where speech and text data are collected with explicit consent and data-sovereignty agreements; (2) AI-assisted documentation and corpus-building, which involves transcription, annotation, and translation; (3) tool and resource development, including dictionary creation and ASR system design, as well as pedagogical application development; (4) community adoption and domain expansion, where tools are introduced to the revitalization landscape, including in education, media, and daily communication; and (5) feedback and refinement, where usage data and community evaluation contribute to the iterative improvement of the AI system and the revitalization strategy as a whole. The model is a cycle, not a line, because Fishman (1991) has emphasized that revitalization is a recursive, ongoing process rather than a one-shot technological solution.

The Power of Information Technology

Rigorous, ethically sound data collection, like Te Hiku Media's tereo Māori system, is unmissable for the high accuracy of the transcription, which is essential for the development of projects that rely on this data for their success and sustainability. The high level of accuracy in the transcription of projects such as Te Hiku Media's tereo Māori system goes hand in hand with rigorous, ethical data collection practices, which are crucial to the success and sustainability of projects that depend on this data. This dependence on data is problematic, however, because, by definition, the languages that are most endangered are those that have the least digital documentation or corpora of text or speech, creating a vicious circle in which the languages that need most help from AI models are the ones that they do not do well on. The tribal and minority languages in India suffer from this problem to the extent that programs like BhashaDaan try to solve it by having data donated by the people (Reboot Democracy, 2026; Government of India, Ministry of Electronics and Information Technology, 2025), but the coverage varies from language to language and across the country.



Translation and Multilingual Infrastructure

India's Bhashini platform is a good example of AI-enabled translation infrastructure at a national scale, using AI speech recognition, machine translation, and text-to-speech technology, and supporting 202 languages, including 22 Scheduled Languages and increasing tribal languages like Adi-Vaani (Digital India BHASHINI Division, 2026). Though this is a big step towards multilingual digital inclusion, the literature indicates a continued mismatch between the extent of language coverage advertised by these platforms and the depth of linguistic correctness achieved for truly low-resource languages, especially those with non-standardized orthographies (AI4Bharat, 2024).

Pedagogical Innovation versus Quality Control

The integration of AI-powered conversational agents into language-learning tools has greatly increased access to the teaching of endangered languages, as in Duolingo's Hawaiian and Navajo courses (Amazon Web Services, n.d.). However, the tensions inherent in the literature were brought into sharp focus by the controversy around Duolingo's "AI-first" pivot, in which human language experts were more or less replaced by generative AI content production – especially for languages with a limited number of authoritative digital resources available to support the outputs of their models (Toppo, 2025).

Frontier Applications

At the extreme end of language endangerment, recent research shows that large language models can also aid revitalization efforts for almost extinct written languages for which little data exists. The NüshuRescue project has achieved high translation accuracy for the critically endangered Nüshu script, historically used by Yao women in China, by fine-tuning a large language model (LLM) with only 35 example sentences (Yang et al., 2024). The present case study suggests that few-shot and transfer learning methods can provide a promising alternative for languages for which it is not otherwise practical to build a corpus, due to a lack of native speakers.

The central ethical fault line is Data Sovereignty (5.6)

In almost all the cases examined, the issue of who should control the language data becomes the main ethical issue. In contrast, there are examples of commercial companies attempting to access Māori voice data without Māori consent for their own commercial products or translation purposes (International Telecommunication Union, 2022). The CARE Principles for Indigenous Data Governance, which stands for Collective Benefit, Authority to Control, Responsibility, and Ethics, offer a structured approach to analysing cases of language-technology use to the benefit of Indigenous communities (Carroll et al., 2020), and Cooper et al. (2024) empirically show that the process of community engagement, not the sophistication of technology, is the most important factor in determining whether Indigenous communities perceive a language-technology project as legitimate.

Table 1

AI Applications across the Stages of Endangered Language Documentation and Revitalization

Stage	AI Technology	Illustrative Project	Reported Outcome
Documentation & Transcription	Automatic Speech Recognition (ASR)	Te Hiku Media, tereo Māori (NVIDIA, 2024)	~92% transcription accuracy; reduced reliance on manual transcription
Corpus & Resource Building	Web crawling, language identification, crowd-sourced voice donation	AI4Bharat IndicVoices; BhashaDaan, India (AI4Bharat, 2024; Govt. of India, 2025)	Large multilingual speech datasets spanning 13+ Indian languages
Machine Translation	Neural Machine Translation; fine-tuned multilingual LLMs	Bhashini National Language Translation Mission (Digital India BHASHINI Division, 2026)	Real-time translation across 22 scheduled and several tribal languages
Pedagogy & Transmission	Gamified apps; conversational AI tutors	Duolingo courses for Hawaiian and Navajo (AWS, n.d.)	Expanded learner reach; quality-control concerns under “AI-first” model (Toppo, 2025)



Revival of Critically Endangered Scripts	Few-shot fine-tuning of large language models	NüshuRescue, China (Yang et al., 2024)	48.69% translation accuracy achieved from only 35 training examples
Governance & Ethics	Indigenous-controlled data licensing frameworks	Kaitiakitanga License; CARE/OCAP principles (Carroll et al., 2020)	Community retains authority over the collection, storage, and use of language data

Most of the AI interventions examined are effective at the documentation and corpus-building phases (roughly levels 6 to 8 of Fishman's GIDS, in which there is little or no intergenerational transmission), which create the raw materials for higher-stage interventions (such as school curricula and broadcast media). But, as Fishman warns, AI tools do not necessarily replace the need for foundational aspects of transmission; rather, they serve as infrastructure to be integrated into a wider community-based revitalization program (Mager et al., 2023). From the perspective of Rogers (2003), diffusion lens, successful programs, such as Te Hiku Media, are those that have high compatibility with the values of the community, are licensed by an Indigenous body, and are made visible to the public through radio and deployment of an app, which, under Rogers theory, are factors that lead to prolonged uptake.

Case Studies and Examples

Te Hiku Media and Te Reo Māori

The Māori-led media organization in New Zealand, Te Hiku Media, created an ASR system for tereo Māori using the open-source NVIDIA NeMo toolkit, which accurately transcribed around 92 percent of tereo Māori and 82 percent of bilingual Māori-English speech (NVIDIA, 2024). Decades of recordings have been placed on the organization's digital archives, and apps like Rongo provide users with pronunciation practice on authentic Māori speech, while the broader team has developed speech recognizers, speech synthesis and real-time pronunciation models for the community by applying machine learning (AI for Good, 2022; International Telecommunication Union, 2022). Most importantly, the project governance framework, the Kaitiakitanga License, ensures that the community is the primary user, and that external use of the community's voice data is legally restricted, directly putting into practice principles of



Indigenous data-sovereignty in the context of a working, and technically complex, AI system. This case illustrates how high technical performance and strong community control is possible at the same time, and provides a template for Indian tribal-language initiatives.

Indigenous Language Models in the Americas

Pinhanez et al. (2024) describe a multi-year collaboration with several Indigenous communities in Brazil, in which AI researchers worked with the communities to create spell-checkers, predictive text, and prototype “Indigenous Language Models” for languages with almost no digital representation. The work is unique in its explicit attention to the ethics of data collection from communities that have traditionally been the object of extractive research practices, and in the suggestion that for some critically endangered languages, an interactive language model may serve as a long-lasting way of documenting – a digital repository that future generations may be able to access after the last living fluent speaker.

NushuRescue: AI for a Critically Endangered Script

The NushuRescue project (Yang et al., 2024) demonstrates the promise of LLM in revitalization even with limited data. The researchers then adapted the LLM to work with Nūshu, a script previously used only by women in Jiangyong County, China, and fine-tuned it with just 35 parallel sentence examples, resulting in meaningful translation results and the creation of a larger set of Nūshu and translated sentences. The case is particularly relevant for the country's most endangered languages, which have limited numbers of elders in their community who are still speaking them and few or any digital materials to document them.

Bhashini and India's Multilingual AI Infrastructure

The Bhashini platform launched by the Government of India, under the National Language Translation Mission, is a case in point of a State-centric initiative that involves speech recognition, translation, and TTS across the twenty-two Scheduled Languages and tribal languages, with complementary projects like Adi-Vaani and BharatGen (Digital India BHASHINI Division, 2026; Government of India, Ministry of Electronics and Information Technology, 2025). The BhashaDaan voice donation program directly tackles the issue of data scarcity highlighted in the entire paper by urging the masses to donate their voice in their own

native languages, such as Konkani, Bodo, etc., which are relatively less resourced Scheduled Languages (Reboot Democracy, 2026). Bhashini currently does not cover the entire spectrum of languages in India that are severely endangered, but is an important source of national infrastructure upon which can be developed more targeted and community-specific revitalization projects that learn from the work of Te Hiku Media and other Indigenous-led projects.

Sanskrit and Classical Language Speech Technology

In the case of a classical language with limited vernacular use and a rich textual heritage, the development of an ASR system and a speech corpus for Sanskrit is unique. While critically threatened oral languages have extensive historical texts, Sanskrit has an extensive corpus of historical texts but a scarcity of annotated speech data that reflects contemporary pronunciation practices. This case demonstrates that the notion of “endangerment” in India spans a spectrum, from those that face domain loss to those that face speaker loss, each of which requires a different approach to AI.

Conclusion

This paper examined the expanding role of Artificial Intelligence (AI) in the conservation and revitalization of endangered languages, with particular emphasis on the opportunities and challenges within the Indian context. By analysing diverse initiatives, including the community-led automatic speech recognition (ASR) model developed by Te Hiku Media for tereo Māori, India’s AI-driven language platforms such as Bhashini, and the data-efficient NushuRescue project, the study demonstrates that AI can significantly strengthen language documentation, translation, speech technologies, educational resources, and digital accessibility for low-resource languages. These examples illustrate that AI has the potential to enhance the scale, efficiency, and accessibility of language revitalization efforts when implemented in culturally appropriate and community-oriented ways.

The analysis further reveals that technological innovation alone is insufficient to reverse language decline. Successful language revitalization depends on community ownership of linguistic resources, meaningful participation in decision-making, and the integration of AI within existing systems of intergenerational language transmission rather than its substitution.



Interpreted through Fishman's Reversing Language Shift framework and UNESCO's language vitality indicators, the findings reinforce the view that AI should function as an enabling technology that complements community knowledge, cultural practices, and language planning. In this context, the proposed AI-Mediated Revitalization Cycle provides a conceptual framework for understanding how technological innovation can support sustainable language revitalization while respecting cultural autonomy and linguistic rights.

A key contribution of this study is its interdisciplinary perspective, which integrates advances in artificial intelligence with sociolinguistic theory, language policy, and ethical governance. By focusing on India, one of the world's most linguistically diverse countries and an emerging leader in AI-enabled language technologies, the study addresses a significant gap in existing scholarship. India's expanding Digital Public Infrastructure and national initiatives such as Bhashini and BhashaDaan create unprecedented opportunities for documenting and revitalizing endangered languages. However, these initiatives should move beyond data collection alone by promoting community participation, equitable data governance, linguistic inclusion, and long-term institutional support, drawing lessons from successful community-led revitalization efforts across the world.

Future research should adopt stronger interdisciplinary collaborations between computer scientists, linguists, anthropologists, and language communities to develop inclusive AI systems for low-resource languages. Particular attention should be given to creating robust NLP benchmarks for India's tribal and minority languages, improving multilingual speech and translation technologies, evaluating the long-term social and educational impacts of AI-assisted revitalization, and establishing ethical frameworks that ensure transparency, fairness, and community ownership of linguistic data. Comparative studies across multilingual societies would also enhance understanding of how AI can be adapted to different cultural and institutional contexts.

Artificial Intelligence is not a substitute for the social, cultural, and intergenerational processes that sustain endangered languages. Nevertheless, when developed through participatory approaches, guided by ethical principles, and aligned with community aspirations,



AI can become a transformative instrument for preserving linguistic diversity and strengthening cultural heritage. The future of endangered language revitalization will therefore depend not only on advances in artificial intelligence but also on sustained collaboration between technology developers, policymakers, researchers, and language communities to ensure that digital innovation contributes to an inclusive and linguistically diverse future.

References

- Adiga, D., Kumar, R., Krishna, A., Jyothi, P., Ramakrishnan, G., & Goyal, P. (2021). Automatic speech recognition in Sanskrit: A new speech corpus and modeling insights. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (pp. 5039–5050). Association for Computational Linguistics.
- AI for Good. (2022, July 26). Accelerating the revitalization of tereo Māori with AI. International Telecommunication Union. <https://aiforgood.itu.int/event/accelerating-the-revitalisation-of-te-reo-maori-with-ai/>
- AI4Bharat. (2024). IndicVoices: Towards building an inclusive multilingual speech dataset for Indian languages (arXiv:2403.01926). arXiv. <https://arxiv.org/abs/2403.01926>
- Amazon Web Services. (n.d.). AI helps Duolingo personalize language learning. AWS Machine Learning Customer Stories. Retrieved June 19, 2026, from <https://aws.amazon.com/machine-learning/customers/innovators/duolingo/>
- Bird, S. (2020). Decolonising speech and language technology. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 3504–3519). International Committee on Computational Linguistics.
- Bromham, L., Dinnage, R., Skirgård, H., Ritchie, A., Cardillo, M., Meakins, F., Greenhill, S., & Hua, X. (2022). Global predictors of language endangerment and the future of linguistic diversity. *Nature Ecology & Evolution*, 6(2), 163–173. <https://doi.org/10.1038/s41559-021-01604-y>



- Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE principles for Indigenous data governance. *Data Science Journal*, 19(1), 43. <https://doi.org/10.5334/dsj-2020-043>
- Cooper, N., Heldreth, C., & Hutchinson, B. (2024). “It’s how you do things that matters”: Attending to process to better serve Indigenous communities with language technologies. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 204–211). Association for Computational Linguistics.
- Cultural Survival. (2025, August 8). Indigenous peoples and AI: Defending rights, shaping the future of technology. <https://www.culturalsurvival.org/news/indigenous-peoples-and-ai-defending-rights-shaping-future-technology>
- Digital India BHASHINI Division. (2026, March 6). BHASHINI: Bridging 1.4 billion citizens to services and opportunities via language AI. EletseGov. <https://egov.eletsonline.com/2026/03/bhashini-bridging-1-4-billion-citizens-to-services-and-opportunities-via-language-ai/>
- Fishman, J. A. (1991). Reversing language shift: Theoretical and empirical foundations of assistance to threatened languages. *Multilingual Matters*.
- Government of India, Ministry of Electronics and Information Technology. (2025, October 25). India’s 22 languages go digital: AI platforms like Bhashini, BharatGen and Adi-Vaani lead the multilingual revolution. DD News. <https://ddnews.gov.in/en/indias-22-languages-go-digital-ai-platforms-like-bhashini-bharatgen-and-adi-vaani-lead-the-multilingual-revolution/>
- Haokip, T. (2022). Artificial intelligence and endangered languages. *Journal of North East India Studies*, 12(1), 1–7.



- Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2021). A survey on recent approaches for natural language processing in low-resource scenarios. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 2545–2568). Association for Computational Linguistics.
- Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In Proceedings of the 37th International Conference on Machine Learning (pp. 4411–4421). PMLR.
- International Telecommunication Union. (2022, August 9). How AI is helping revitalise indigenous languages. ITU Hub. <https://www.itu.int/hub/2022/08/ai-indigenous-languages-maori-te-reo/>
- Inuwa-Dutse, I. (2025). NaijaNLP: A survey of Nigerian low-resource languages (arXiv:2502.19784). arXiv. <https://arxiv.org/abs/2502.19784>
- Kavitha, R., et al. (2023). AI: A catalyst for language preservation in the digital era. Shanlax International Journal of English, 12(S1), 102–106.
- Mager, M., Oncevay, A., Rios, A., Mirzakhlov, J., & Kann, K. (2023). The role of computational linguistics in indigenous language revitalization: Challenges and opportunities. In Proceedings of the 1st Workshop on Computation for Indigenous Languages (C3NLP) (pp. 23–31).
- NVIDIA. (2024, November 5). Māori speech AI model helps preserve and promote New Zealand Indigenous language. NVIDIA Blog. <https://blogs.nvidia.com/blog/te-hiku-media-maori-speech-ai/>
- Pinhanez, C., Cavalin, P., Storto, L., Finbow, T., Cobbinah, A., Nogima, J., Vasconcelos, M., Domingues, P., Mizukami, P. de S., Grell, N., Gongora, M., & Gonçalves, I. (2024).



Harnessing the power of artificial intelligence to vitalize endangered Indigenous languages: Technologies and experiences (arXiv:2407.12620). arXiv. <https://arxiv.org/abs/2407.12620>

Reboot Democracy. (2026, January 21). Voices in every language: How India is building more inclusive AI. <https://rebootdemocracy.ai/blog/voices-in-every-language-bhashadaan-india>

Roberts, J. S., & Montoya, L. N. (2023). In consideration of Indigenous data sovereignty: Data mining as a colonial practice (arXiv:2309.10215). arXiv. <https://arxiv.org/abs/2309.10215>

Rogers, E. M. (2003). Diffusion of innovations (5th ed.). Free Press.

Toppo, G. (2025, June 3). As Duolingo turns to AI, some users say language app has joined “the dark side.” The 74. <https://www.the74million.org/article/as-duolingo-turns-to-ai-some-users-say-language-app-has-joined-the-dark-side/>

UNESCO Ad Hoc Expert Group on Endangered Languages. (2003). Language vitality and endangerment. UNESCO. <https://ich.unesco.org/doc/src/00120-EN.pdf>

Visual Capitalist. (2024, February 8). The state of the world’s 7,168 living languages. <https://www.visualcapitalist.com/cp/state-of-the-worlds-living-languages/>

Yang, I., Ma, W., & Vosoughi, S. (2024). NushuRescue: Revitalization of the endangered Nüshu language with AI (arXiv:2412.00218). arXiv. <https://arxiv.org/abs/2412.00218>