International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 04, April 2026)

Making India's Tribal Languages Searchable: AI-Driven Multilingual Platforms for Preservation, Access, and Governance

Rajib Kumar Guin
Librarian, Silda Chandrasekhar College, Affiliated to Vidyasagar University, Midnapore, West Bengal, India

***Abstract*--The 461 native tribal languages of India, including Santali, Gondi, Bhili, Mundari, and many others, are under severe threat because they have few written records, are oral languages, and are excluded from mainstream digital literature. Neural machine translation (NMT), multilingual embeddings, and speech technologies are now used together to make these languages both text- and query-readable, as well as speech- and image-queryable, in AI-driven multilingual search systems, including Bhashini, BharatGen, and Adi-Vaani. Multilingual embeddings and retrieval-augmented generation (RAG) workflows can expand access while raising critical governance issues. The paper will explore how these platforms can close digital gaps in tribal areas, enable semantic access to cultural knowledge, facilitate schemes such as PM JANMAN, and address data scarcity, dialect variability, and data sovereignty. We believe that a well-managed AI pipeline can uphold language traditions while also providing space for local innovation in the future Indian AI landscape.**

***Keywords*-- Adi-Vaani, BharatGen, Bhashini, Neural Machine Translation, RAG**

I. Introduction

India's tribal communities constitute over 10% of the national population and collectively speak around 461 tribal languages, many of which lack written traditions and digital presence [6]. UNESCO and national surveys identify over 100 of these as vulnerable or endangered, with declining intergenerational transmission and a rapid shift to dominant languages such as Hindi and regional state languages [7]. The resulting "digital language divide" is stark: mainstream search and online services have historically been optimized for English and a few major Indian languages, marginalizing Adibasi users and rendering their knowledge systems largely invisible to standard information retrieval [11]. Recent government-led initiatives under Digital India and PM JANMAN embed AI-based translation and speech technologies into public digital infrastructure, seeking to provide information and services in tribal mother tongues.

Against this backdrop, multilingual AI search platforms promise not only improved access but also the systematic documentation and preservation of endangered languages through everyday use [2].

II. India's Indigenous Linguistic Landscape

India's tribal linguistic diversity spans several major families, including Austro-Asiatic (e.g., Munda languages such as Santali and Mundari), Dravidian (e.g., Gondi), and Indo-Aryan branches (e.g., Bhili), each with distinct phonology, morphology, and script practices. Many of these languages feature rich agglutinative or polysynthetic morphology and extensive dialect continua, which complicate tokenization, normalization, and lexicon design for NLP models [8]. Census linked and academic surveys indicate that, among 461 tribal languages, 81 are classified as vulnerable and 42 as critically endangered, often with speaker populations under 10,000 and limited literacy in the mother tongue. Santali alone has over seven million speakers, yet many other Adibasi languages survive primarily through oral traditions, making them heavily dependent on audio recordings, folklore, and community memory rather than print. Documentation programmes such as the Scheme for Protection and Preservation of Endangered Languages (SPPEL) and CIIL's Sanchika have begun to assemble lexical resources, narratives, and primers that now serve as crucial training data for AI models [5].

III. AI Foundations For Multilingual Search

3.1 Neural Machine Translation (NMT)

Modern multilingual search for tribal languages relies heavily on neural machine translation models such as No Language Left Behind (NLLB-200) and IndicTrans2, which are fine-tuned on low-resource corpora to support bidirectional translation between Hindi, English, and target tribal languages [9].

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 04, April 2026)

These systems enable cross lingual search workflows in which user queries in a high resource language are translated into the tribal language for document retrieval, and the retrieved content is then translated back for display. Platforms like Adi-Vaani and NLLB, and IndicTrans2, are specifically adapted for six Adibasi languages—Bhili, Mundari, Gondi, Santali, Kui, and Garo—allowing near real-time text-to-text and speech-to-text translation in both directions.

3.2 Multilingual Embeddings and Semantic Retrieval

Beyond surface translation, multilingual language models such as MuRIL and mT5 provide shared embedding spaces that align semantically similar sentences across many Indian languages, including code mixed content [10]. This supports semantic search in which a query expressed in Hindi, English, or a code switched mixture can retrieve relevant Adibasi documents that are paraphrased or expressed in different dialects. Such embedding based retrieval is particularly important for folklore and oral histories, where literal lexical overlap may be low but thematic and narrative similarity is high.

3.3 Speech Technologies: ASR and TTS

Automatic speech recognition (ASR) and text-to-speech (TTS) systems developed under initiatives like AI4Bharat and BharatGen extend multilingual search to voice-first interfaces, which are often more natural for low-literacy users. Using corpora from SPPEL and other field recordings, these models can process dialectal variation and background noise to support voice queries in tribal languages and to read out retrieved content in a familiar accent. In practical deployments, tribal users can, for example, speak a query in Gondi to access health advisories or welfare scheme details, with responses synthesized back into Gondi or another preferred language.

3.4 Retrieval Augmented Generation (RAG)

Retrieval-augmented generation pipelines combine vector search over multilingual embeddings with large language models (LLMs) trained on Indic data to produce contextualized, grounded answers. In the tribal-language setting, RAG enables systems to pull from curate folklore archives, government circulars, and educational materials, and then generate concise explanations or translations into the user's language while citing the underlying sources. This reduces the risk of hallucinations and aligns AI outputs with verified cultural and administrative content, which is critical for sensitive domains such as welfare delivery and heritage documentation.

IV. GOVERNMENT-LED MULTILINGUAL PLATFORMS

4.1 Adi Vaani

Adi Vaani is India's first dedicated AI-powered translator for tribal languages, launched in beta by the Ministry of Tribal Affairs as part of Janjatiya Gaurav Varsh. Accessible via web and mobile applications, the platform currently supports translation between Hindi, English, and six Adibasi languages (Bhili, Mundari, Gondi, Santali, Kui, and Garo), with further expansion planned. Architecturally, Adi-Vaani builds on NLLB and an IndicTrans2 model adapted for low-resource settings, offers text-to-text, speech-to-text, text-to-speech, and speech-to-speech translation, and integrates OCR to digitize manuscripts, primers, and archival documents [1]. The system is being tested through programmes such as Adi Karmayogi, which targets capacity building for volunteers across over 100,000 villages, ensuring that the models improve through community feedback while directly serving local governance needs.

4.2 Bhashini

Bhashini, implemented by the Digital India BHASHINI Division under the Ministry of Electronics and IT, serves as the national language translation mission platform for 22+ Indian languages and several dialects. It offers APIs for text-to-text translation, ASR, TTS, OCR, and language detection, enabling developers and government departments to embed multilingual capabilities into websites, applications, and IVR systems. Bhashini has been integrated with platforms such as e-Shram, e-Gram Swaraj, CPGRAMS, AICTE, and UGC portals, allowing beneficiaries, students, and citizens to access content and services in their preferred languages. For tribal languages, Bhashini's embedding and translation services facilitate cross lingual search across scripts and dialects, making Adibasi content discoverable alongside mainstream Indian languages [2].

4.3 BharatGen and AI4Bharat

BharatGen functions as a sovereign AI stack that provides open-source models and datasets for Indian-language understanding, generation, and speech processing, including modules tailored for tribal and endangered languages. By leveraging SPPEL corpora and CIIL's Sanchika repository, BharatGen supplies TTS and STT models that form the backbone of voice based search interfaces in low resource languages.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 04, April 2026)

AI4Bharat, an academic-- industry consortium, releases open-source NMT, ASR, and embedding models that can be fine-tuned to specific Adibasi languages, thereby lowering entry barriers for startups and community projects building specialized retrievers [3][4].

V. Multilingual Search Workflows And Applications

Multilingual search platforms typically pipeline several AI components: user queries are accepted in text or speech, normalized and tokenized, optionally translated via NMT, embedded into a multilingual semantic space, matched against indexed documents or transcripts, and finally presented to the user or read aloud in the user's chosen language. For example, a Hindi query about agricultural practices can be translated into Santali, used to search Santali-language guides and oral histories, and the relevant results summarized back into Hindi or Santali. OCR capabilities enable the ingestion of printed primers and handwritten folklore notebooks into digital corpora, while speech-to-text pipelines convert recorded oral narratives into searchable text linked to the original audio.

These capabilities support several concrete applications in tribal contexts:

Governance: Delivery of scheme information, grievance redressed forms, and official advisories under PM JANMAN and related programmes in local Adibasi tongues.

Education: Digital textbooks, storybooks, and primers in tribal languages, aligned with NEP 2020's emphasis on mother tongue instruction and hosted on platforms like SWAYAM and e-KUMBH.

Cultural preservation: Searchable repositories of songs, myths, and rituals, where users can query by theme, location, or clan and retrieve narratives in their original language.

In many cases, AI-enabled pipelines have accelerated documentation: estimates indicate that automated transcription and translation can scale language archiving an order of magnitude faster than purely manual workflows.

VI. Impact On Tribal Communities: Case Snapshots

Early evidence from pilot deployments suggests substantial gains in both access and preservation.

Adi-Vaani has been used to document folklore and daily communication in tribal districts, with some pilots reporting up to 30% improvement in effective access to welfare information when interfaces are localized into native languages. Tribal Research Institutes in states such as Odisha and Jharkhand now employ AI tools to create and maintain digital archives that can be searched nationally, connecting local knowledge with researchers, educators, and policymakers. Datasets like ELR-1000 enable evaluation of large language models on culturally grounded tasks for Eastern tribal languages, improving the quality of semantic retrieval and generation in real deployments [13].

At the ecosystem level, startups such as Sarvam AI and independent developers are leveraging open models to build Indic search engines and assistants that outperform generic global systems at handling local nuances, including code-switched queries and region-specific terminology. Such innovation helps ensure that tribal languages are not confined to specialized portals but are integrated into broader AI products used by both mainstream and Adibasi users [12].

VII. Key Challenges

Despite promising progress, several structural and technical challenges remain. First, data scarcity is acute for many tribal languages: limited written material, few recorded speakers, and fragmented orthographic conventions mean that training corpora are often too small for robust NMT and ASR. Synthetic data generation through back translation, data augmentation, and cross lingual transfer can provide partial relief but must be carefully validated to avoid propagating errors or erasing dialectal richness. Second, dialect diversity—for example, among Mundari sub-varieties—makes it difficult to design a single model that generalizes across communities without privileging one dialect over others.

Ethical concerns are equally significant. Questions of data sovereignty and consent arise when community narratives, songs, and ritual knowledge are digitized and used to train commercial or state-run models. Without explicit, ongoing community involvement, such efforts risk replicating extractive patterns in which indigenous knowledge is appropriated without equitable benefit sharing or control. Infrastructure gaps—such as unreliable connectivity and limited device access in remote tribal areas—further constrain real world use, even when sophisticated AI models exist in principle.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 04, April 2026)

VIII. Future Directions And Policy Recommendations

Looking ahead, integrating multilingual search platforms with national digital public infrastructures like UPI and ONDC could enable commerce, entrepreneurship, and service delivery in tribal languages, not only for information access but also for transactions. Roadmaps for Adi-Vaani envision expansion to over 100 tribal and minor languages by 2027, using crowd sourced datasets and community driven annotation to scale coverage while maintaining cultural fidelity. Hybrid AI–human validation loops, where local speakers review and correct translations, will remain essential to ensuring that the semantics and pragmatics of Adibasi discourse are respected.

From a governance perspective, embedding federated learning and on device training can help protect sensitive linguistic and cultural data, allowing models to improve without centralizing raw community corpora. Regulatory frameworks should codify principles of free, prior, and informed consent for data collection and model training, alongside mechanisms for community oversight and redress. Finally, sustained investment in connectivity, devices, and localized user interfaces will be needed to ensure that AI-enabled multilingual search translates into tangible empowerment for speakers of India's most vulnerable languages.

IX. Conclusion

AI-driven multilingual search platforms mark a decisive shift in how India approaches its indigenous linguistic heritage, moving from sporadic documentation towards living, interactive ecosystems where tribal languages function as full participants in the digital public sphere.

By combining NMT, multilingual embeddings, speech technologies, and RAG pipelines, initiatives like Adi Vaani, Bhashini, and BharatGen demonstrate that even severely under resourced languages can be made searchable, speakable, and teachable at scale—provided that communities remain at the centre of design, governance, and benefit sharing.

REFERENCES

[1] Ministry of Tribal Affairs, Government of India. (2024). *Adi Vaani: AI-powered translator for tribal languages*. Janjatiya Gaurav Varsh Initiative.

[2] Digital India BHASHINI Division, Ministry of Electronics and IT. (2024). *Bhashini: National language translation mission platform*. Government of India.

[3] AI4Bharat. (2023). *Open-source neural machine translation and speech models for Indian languages*. AI4Bharat Consortium.

[4] BharatGen Project. (2023). *Sovereign AI stack for Indian language understanding and generation*. Government of India.

[5] CIIL (Central Institute of Indian Languages). (2022). *Sanchika: Lexical and narrative resources for tribal languages*. Mysore: CIIL Press.

[6] Scheme for Protection and Preservation of Endangered Languages (SPPEL). (2022). *Lexical and audio corpora for tribal languages*. Ministry of Education, Government of India.

[7] UNESCO. (2021). *Atlas of the World's Languages in Danger: India*. Paris: UNESCO Publishing. Retrieved from [UNESCO Atlas].

[8] No Language Left Behind (NLLB-200). (2023). *Neural machine translation model for low-resource languages*. Meta AI Research.

[9] IndicTrans2. (2023). *Neural machine translation for Indian languages*. AI Research Labs, India.

[10] MuRIL and mT5 Models. (2023). *Multilingual embeddings for Indian language semantic retrieval*. AI4Bharat Consortium.

[11] NEP 2020. (2020). *National Education Policy 2020: Mother-tongue instruction emphasis*. Ministry of Education, Government of India.

[12] Sarvam AI. (2023). *Indic search engines and AI assistants for tribal languages*. Startup ecosystem report..

[13] ELR-1000 Dataset. (2023). *Evaluation corpus for culturally grounded tasks in Eastern tribal languages*. AI4Bharat Repository.