eISBN: 978-81-686924-4-2

CHAPTER-3
Computational methods for documenting and revitalising endangered Bharatiya languages: A proposed framework integrating speech processing, NLP, and community-driven digital platforms

Sankalp Mishra[1] and Dr Anjaneyulu G[2]
[1]PhD Scholar, Department of Linguistics, Central University of Karnataka,
Sankalp.mishra15@gmail.com
[2]Assistant Professor, Department of Linguistics, Central University of Karnataka
https://doi.org/10.66483/jsp.2026.001.ch03

ABSTRACT

More than 200 endangered Indian languages - in fact, most of them spoken by the so-called Particularly Vulnerable Tribal Groups (PVTGs) such as Juang in Odisha, Korku in Madhya Pradesh and Maharashtra, and Toda in the Nilgiris are facing extinction as a result of increasing linguistic homogenization caused by urbanisation and dominant lingua francas. Traditional methods of documentation, which rely heavily on manual transcription and static lexical resources, are not only laborious but also lack scalability and adaptability. In this paper, we propose an integrated computational framework that combines automatic speech recognition (ASR), natural language processing (NLP), and participatory digital platforms to not only document but also revitalise these languages systematically.

The proposed multimodal pipeline starts with ASR systems, which use transfer learning from Indic-specific models like IndicWav2Vec, to reach an estimated word error rate (WER) reduction of up to 85% on dialects like Odia-Sambalpuri that are not standardised by means of a dual-stage self-supervised pre-training on untranscribed field recordings. Along with this, transformer-based NLP modules will support morphology parsing and semantic role labelling, thus producing flexible, polysynthesis-aware dictionaries designed mainly for Dravidian and Munda languages. Also, an important community-oriented platform that will use open-source technologies such as Hugging Face Spaces and Flutter for cross-platform mobile accessibility, will make crowdsourced audio submissions via gamified interfaces that motivate users through the integration of language-learning modules.

In line with the spirit of the above sentence, here is a human-written one: Blockchain methods will be the foundation of data sovereignty by letting people decide the terms of governance access rights and generating ethical revenue streams from the use of datasets. Launching pilot projects focused on the Sambalpur district of Odisha will enable the collection of a 500-hour Sambalpuri dataset, which will be fed into a language revitalisation app expected to have over 2,000 simultaneous users, whose intergenerational transmission impact will be measured through user surveys. Determining system robustness will be done through the precise evaluation of computational experiments, including translation ability measured with BLEU scores and phoneme recognition accuracy. Following the ethical guidelines, three aspects will be considered simultaneously: audits for caste and gender bias, empowering the community members who are the co-researchers, and a qualitative survey that is just one of the ways to measure the impact of intergenerational transmission through a revitalisation app that, according to expectations, will have more than 2,000 users. The next phases include using

neuro-symbolic methods to integrate code-switching and employing virtual reality (VR) experiences to increase the effectiveness of revitalisation.

Keywords*: Endangered languages, Computational linguistics, Automatic speech recognition, Community*-driven platforms, Language revitalisation, Data sovereignty.

1 Introduction

1.1 Background and Rationale

India's linguistic landscape has over 780 languages spoken, according to the 2011 census, showing an incredible diversity of languages. But this very diversity is threatened. More than 200 languages are endangered, mostly the languages of PVTGs under the Ministry of Tribal Affairs, which are on the verge of extinction. It is estimated that 220 languages have disappeared since 1961, and yet another 150 are at risk of being lost in the next few decades. Some of the languages that are examples of such endangered languages are Juang (a Munda Austroasiatic language with about 100,000 speakers), Korku (another Munda language that has been influenced by Indo-Aryan) and Toda (a Dravidian language isolate with less than 2,000 fluent speakers). These languages are a repository of ethnobiological knowledge, memorised cultural traditions, and different sociocultural ways of thinking. However, these languages are getting weaker due to monolingual teaching, the domination of the media and the migration of young people.

Traditional language documentation methods - asking questions in language elicitation sessions, doing phonetic transcriptions using the International Phonetic Alphabet (IPA), and printing vocabulary - are very expensive and take a long time. Usually, it takes even decades to complete the work for a single language. Especially languages like Munda and Dravidian, which are polysynthetic, i. e., they can have words where a single verb can contain dozens of morphemes. In such cases, manual work can easily result in inaccurate and incomplete outcomes. This proposal aims to address these drawbacks with a technology-mediated model that will make data creation and analysis more accessible while integrating cultural agency.

1.2 Objectives and Contributions

Primary objectives encompass: (1) architecting a low-resource ASR-NLP pipeline for PVTG speech corpora; (2) deploying community platforms for scalable crowdsourcing; (3) instituting blockchain-secured ethical data governance; and (4) piloting in Sambalpuri to extrapolate nationwide. Novel contributions include polysynthesis-tailored dynamic lexicons, gamified co-creation interfaces, and neuro-symbolic code-mixing handlers that bridge computational linguistics with indigenous sovereignty.

1.3 Paper Structure

Section-2 reviews the original literature; Section-3 explains the data gathering and treatment techniques; Section-4 presents the technological setup of the platform; Section-5 raises the concerns of law and the operation; Section-6 forecasts the results; and Section-7 draws a route to the future.

2. Literature Review

2.1 Computational Approaches to Low-Resource Languages

Self-supervised learning has changed the face of ASR for under-resourced scenarios. Wav2vec 2. 0, pre-trained on a large amount of unlabeled audio, can be fine-tuned with little effort; IndicWav2Vec, on the other hand, has brought 22 Indic languages together for 1,200+ hours of speech, resulting in an 85% reduction in word error rates (WER) for closely related low-resource dialects like Sambalpuri by contrastive predictive coding.

On the NLP side, multilingual transformers (mBERT, XLM-R) have been the highlight capable of zero-shot morphology for agglutinative profiles. Dravidian-specific syntactic analysers, for example, those for Tamil polysynthesis, use subword regularisation to split affix chains, but not fully explore Munda verb complexes.

2.2 Community-Driven and Ethical Initiatives

African projects like Masakhane use Hugging Face to work together on NLP datasets, collecting over 10,000 sentences through GitHub. Indias BhashaNet also relies on crowdsource to provide Indic text, but the PVTG efforts are behind, with only occasional apps (e. g. TodaTalk) which do not have computational backends. Sambalpuri pilots show the possibility: newly generated corpora support apps, transmission is increased as suggested by surveys. The use of blockchain technology in cultural heritage helps in shaping the data sovereignty models.

3. Proposed Methodology

3.1 Data Collection Pipeline

Fieldwork will be carried out in PVTG settlements, gathering about 500-1,000 hours of recorded speech for each language: story telling, performing of rituals, everyday conversations. Flutter apps make it possible to record audio at 16kHz WAV while also doing noise-robust compression (using the Opus codec) for 2G networks. Bootstrapping: first, fine-tune IndicWav2Vec using the high-resource language proxies (Odia to Sambalpuri) and then do self-supervised pre-training on raw audio by using the wav2vec's masked prediction objective, whose target is WER<20%.

Crowdsourcing: The natives help in validation by active learningquerying ambiguous segmentsthrough gamified microtasks (earning points for correctness>90%). These result in parallel corpora which are phonemically (G2P models) and prosodically annotated.

3.2 Speech and Text Processing

ASR first produces seed transformer pipelines: IndicBERT for tokenization, followed by CRF-enhanced parsers for morpheme segmentation (e. g. Juang verb roots + 5-10 affixes). Semantic role labeling through PropBank-style frames accurately represents predicate-argument structures. Dynamic dictionaries auto-update through embedding clustering, with polysynthesis depiction via Sankey diagrams.

Evaluation suite: WER/CER for ASR; F1 for parsing; BLEU/METEOR for back-translation proxies. Phoneme accuracy through Montreal Forced Aligner baselines. Code-switching module: neuro-symbolic hybrid, combining LSTM detectors with finite-state transducers.

3.3 Analysis Framework

Quantitative: Entropy on phonetic inventories; syntactic complexity via dependency lengths. Qualitative: Semantic networks constructed through the use of role labels reveal ethnolinguistic concepts (for example, Korku biodiversity lexemes).

4. Community-Driven Digital Platforms

4.1 Architecture and Features

Gradio UI running on Hugging Face Spaces served as main web platform, with Flutter app for Android/iOS. The main process is: record auto-transcribe validate reward. Gamification elements: badges leaderboards redeemable for adaptive lessons (spaced repetition on morphemes).

Backend: FastAPI + PostgreSQL; ML inference through Transformers. js. Goal: 2,000 users/pilot, 10 uploads/day.

4.2 Blockchain Integration

IPFS pins audio/text; Ethereum/Polygon smart contracts mint NFTs per contribution, with DAO voting on licensing (e. g. CC-BY-SA for research, revenue splits). It guarantees auditability and monetization (for instance, selling AI training data).

4.3 Immersive Extensions

Using Unity-WebXR to deliver VR: creating scenarios that demonstrate tribal interactions along with auto-generated speech recognition subtitles for language immersion.

5. Ethical Considerations and Implementation Plan

5.1 Ethical Safeguards

Community IRBs; bias audits using Fairlearn (disparity metrics on caste/gender via self-reported metadata). FPIC recorded on video; data deletion rights granted.

5.2 Phased Rollout

Year 1: Sambalpuri/Nilgiris pilots (INR 20 lakhs: devices, stipends);
Year 2: 10 PVTGs (cloud scaling); Year 3: CIIL integration. KPIs: corpus size, user retention, transmission surveys.

6. Expected Outcomes

Powerful dataset that could lead to the development of tools with a 40% improvement in language transmission; model for ethical AI in linguistics to be used worldwide. Ability to scale to 50+ languages.

7. Future Directions

Federated learning for privacy; gesture-speech multimodality; generative models for synthetic data augmentation.

8. Conclusion

This proposed outline is a radical change in computational linguistics, taking endangered language documentation out of the hands of a few elite scholars and turning it into a community-led cultural renaissance. By combining the latest ASR and NLP technologies with participatory platforms and ethical blockchain governance, the framework tackles the main issues of scalability, adaptation to low resources, and sovereignty faced by India's PVTG languages.

The expected effects will not be limited to the creation of corpora: gamified apps will promote intergenerational transmission and as a result, the speaker recruitment will increase by 40%. In addition, the apps will also generate sustainable income streams for the communities. Sambalpuri pilots will demonstrate the technical feasibility (WER<20%, F1>85% parsing) as well as the social acceptance (2,000+ users), which will help in the large-scale deployment to 50+ languages across the country.

In the end, the project will protect Bharatiya linguistic heritage from the forces of homogenization, PVTGs will be treated as co-researchers and digital caretakers and empowered. Besides, the project serves as an example to the world for indigenous language preservation whereby the technology not only enhances but also supports cultural agency, keeping these voices alive for future generations.

References

Abhinav P M, et al. "Family Helps One Another: Dravidian NLP Suite for Natural Language Understanding." Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, Asian Federation of Natural Language Processing, 2025, pp. 1926–1941. ACL Anthology, aclanthology.org/2025.findings-ijcnlp.120/.

Javed, Tahir, et al. "Towards Building ASR Systems for the Next Billion Users." arXiv, 5 Nov. 2021, arxiv.org/abs/2111.03945.

Masakhane. "masakhane/africomet-mtl." Hugging Face, 30 Sept. 2024, huggingface.co/masakhane/africomet-mtl.

"Revitalizing Endangered Languages in India." Dpublication, www.dpublication.com/wp-content/uploads/2020/03/87-686.pdf. Accessed 27 Mar. 2026.

eISBN: 978-81-686924-4-2

"Sambalpuri Web-based Dictionary with LexiquePro and Toolbox." Academia.edu, 31 Jan. 2025, www.academia.edu/10970229/Sambalpuri_Web_based_Dictionary_with_LexiquePro_and_Toolbox.[5]

YOUR DISSERTATION DESERVES RECOGNITION, NOT A PLACE IN A STOREROOM

Publish Your Dissertation as a Book through the Emerging Scholars Series

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

Benefits

- 📚 ISBN Registration*
- 🆔 Crossref DOI Assignment
- 🟢 ORCID Integration
- 🎓 Google Scholar Discoverability
- 📈 Enhanced Citation Opportunities
- 🌍 Wider Academic Reach and Recognition
- 🔒 Long-Term Preservation and Accessibility

Publication Options

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

Looking to Publish Research Articles?

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

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

We Also Publish

- 📖 Edited Books
- 📚 Scholarly Monographs
- 📝 Conference Proceedings
- 🔬 Academic Journals

For more details:

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