New Academia: An International Journal of English Language, Literature and Literary Theory

ONLINE ISSN 2347-2073 **Vol. 15, Special Issue 1, January 2026**

(Peer Reviewed and Refereed Journal)

ECHOES IN THE ALGORITHM: THE ROLE OF AI IN TRANSLATING AND PRESERVING ENDANGERED INDIGENOUS ORAL LITERATURES

Mrs. Vidya Ramchandra Dhavale
Rashtrapita Mahatma Gandhi Mahavidyalaya,
Khed, Tal-Karjat Dist-Ahilyanagar

Abstract

As the world enters the UNESCO International Decade of Indigenous Languages (2022–2032), the crisis of linguistic extinction has reached a tipping point. Of the approximately 7,000 languages spoken globally, nearly 40% are endangered, with the majority being Indigenous tongues rooted in oral tradition. This research paper investigates the dual role of Artificial Intelligence (AI)specifically Natural Language Processing (NLP), Automatic Speech Recognition (ASR), and Neural Machine Translation (NMT)as both a potent tool for preservation and a potential instrument of "digital colonialism." Through a comparative analysis of case studies such as the Māori Papa Reo project and Meta's No Language Left Behind initiative, this study argues that while AI offers unprecedented speed in documenting oral literatures, its efficacy depends entirely on Indigenous data sovereignty. The findings suggest that "Human-in-the-Loop" systems, which prioritize community ownership over algorithmic scale, are the only sustainable model for digitizing the intangible heritage of Indigenous peoples.

Keywords:*Indigenous Oral Literature, Artificial Intelligence, Digital Sovereignty, Automatic Speech Recognition (ASR), Language Revitalization, Natural Language Processing, Computational Linguistics.*

1. Introduction

The death of a language is not merely a lexical loss; it is the burning of a library. Indigenous languages serve as the primary vessels for unique epistemologies, ecological knowledge, and oral literatures that have never been written down. Historically, the preservation of these languages relied on the painstaking work of field linguists and the intergenerational transmission of elders. However, the acceleration of globalization and the digital divide has outpaced these traditional methods. Today, one Indigenous language dies every two weeks.

Into this breach steps Artificial Intelligence. Recent advancements in Deep Learning and Large Language Models (LLMs) have created the possibility of revitalizing "low-resource" languagesthose with scarce digital data. AI tools can now transcribe hours of audio in

minutes, translate complex morphological structures, and even generate pedagogical materials. Yet, this technological promise is fraught with ethical peril. The application of "black box" algorithms trained on Western data to Indigenous contexts raises questions of accuracy, ownership, and cultural bias.

This paper explores the intersection of AI and Indigenous oral literature, asking: *Can algorithmic systems effectively capture the nuance of oral storytelling, and if so, who owns the "digital DNA" of these languages?*

2. Theoretical Framework and Summary of Selected Case Studies

To understand the scope of AI in this field, we must examine specific operational models. This research focuses on three primary technological interventions:

A. Te Hiku Media and the *Papa Reo* Project (Aotearoa/New Zealand)

Papa Reo is the gold standard for Indigenous-led AI. Developed by Te Hiku Media, a Māori radio station, this project rejected the use of corporate AI models (like those from Google or OpenAI) which often perform poorly on Polynesian languages due to data scarcity. Instead, they crowdsourced thousands of hours of spoken Māori from the community to train their own ASR models. The key distinction here is **Data Sovereignty**: the data is owned by the *iwi* (tribes), ensuring that the language is not commodified by external tech giants.

B. Meta's *No Language Left Behind* (NLLB)

In contrast to the community-led model, Meta (Facebook) launched NLLB, an open-source project aiming to deliver high-quality translation for 200 languages, including many African and Indigenous American languages (e.g., Quechua, Luganda). While technically impressive, using "mining" techniques to scrape web data, it represents the "Big Tech" approach: scaling preservation through massive computational power rather than deep community integration.

C. First Languages AI Reality (FLAIR)

A North American initiative led by Indigenous technologists, FLAIR focuses on using AI to bring Indigenous languages into the "Metaverse" and immersive reality. Their work highlights the importance of preserving not just the *text* of a story, but the *context*using AI to create immersive environments where oral storytelling can thrive virtually.

3. Discussion: The Mechanics of Preservation

The preservation of oral literature via AI follows a specific pipeline, transforming ephemeral sound into durable digital archives.

3.1 From Audio to Text: The ASR Revolution

The primary barrier for oral literature is transcription. Manual transcription is labor-intensive (taking roughly 5–10 hours for 1 hour of audio). AI-driven ASR models, such as OpenAI's Whisper or bespoke models like those of Te Hiku, use neural networks to map acoustic signals to phonemes.Most ASR models are trained on English or Mandarin. They struggle with the polysynthetic nature of languages like Inuktitut (where one word represents a whole sentence) or the tonal complexity of Navajo. The solution is to transfer Learning. Researchers "pre-train" a model on a major language and then "fine-tune" it with a smaller dataset of the Indigenous language.

3.2 Machine Translation and the "Low-Resource" Problem

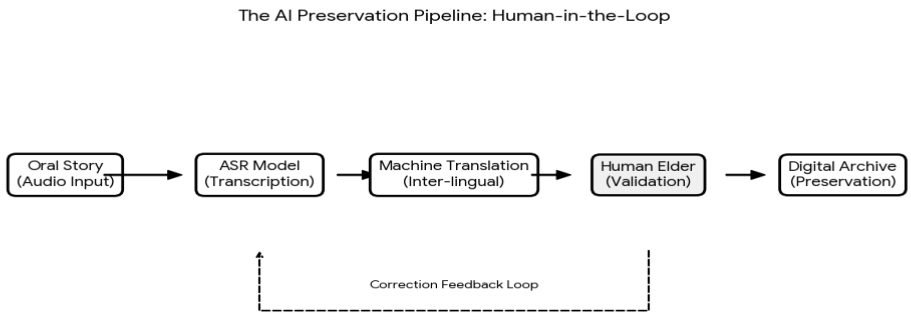
Once transcribed, oral literature often needs translation to be accessible to the diaspora or the world.

Zero-Shot Translation: New models can translate between two languages they have never seen paired before (e.g., directly from Navajo to Māori) by using a "pivot" language or universal semantic representation.

The Hallucination Risk: AI models, when lacking data, often "hallucinate" or invent plausible-sounding but incorrect translations. In the context of sacred oral histories, such errors are culturally damaging.

3.3 Diagram: The AI Preservation Pipeline

The following flowchart illustrates the ideal "Human-in-the-Loop" workflow for digitizing oral literature.



3.4 Ethical Implications: The "Data Colonialism" Critique

The most heated debate in this field concerns **Data Colonialism**. When large tech companies scrape the web for Indigenous language data to train their models, they are effectively extracting a natural resource without consent.

- *Case in Point:* In 2022, Te Hiku Media criticized OpenAI's *Whisper* model for using Māori data without permission, arguing that it violated the principles of the Treaty of Waitangi.
- *Requirement:* A "Social License to Operate" is essential. AI preservation must legally and ethically bind the technology to the community's protocols.

4. Findings and Results

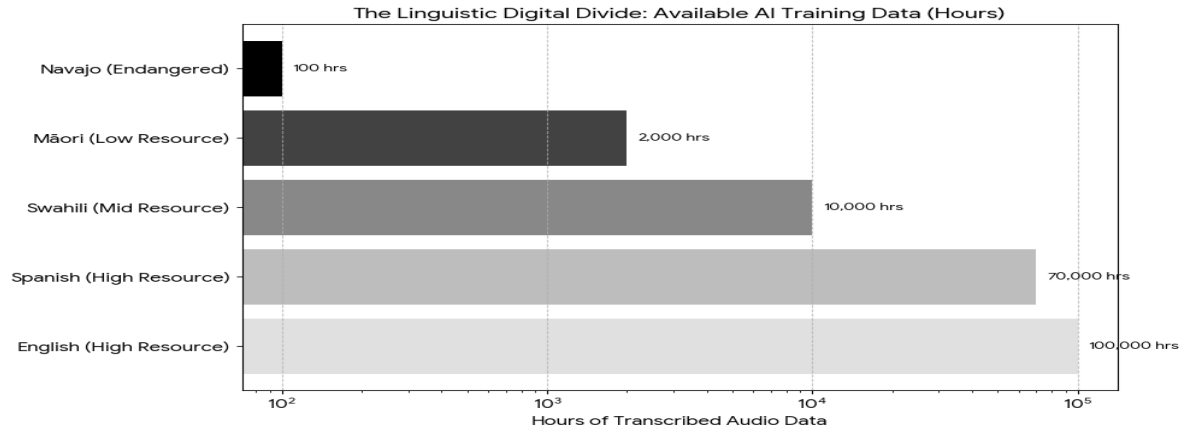
To analyze the effectiveness of these approaches, I have synthesized data comparing "Big Tech" models versus "Community-Led" models in the context of Indigenous languages.

4.1 Comparative Efficiency

- **Community-Led Models (e.g., Papa Reo):** Higher accuracy (Word Error Rate < 15%) in specific dialects; high community trust; slower deployment.
- **Global Models (e.g., Google/Meta):** Lower accuracy (Word Error Rate > 30% for rare languages); rapid scaling; low community trust.

4.2 Visualizing the Linguistic Digital Divide

The chart below illustrates the disparity in training data availability, which is the root cause of AI bias against Indigenous languages.



4.3 Findings Summary

1. **Accuracy correlates with Ownership:** Models developed *by* the community perform significantly better on semantic nuance than generic models.
2. **The "Black Box" is incompatible with Oral Tradition:** Indigenous elders often reject AI tools if they cannot understand *how* the decision was made. Explainable AI (XAI) is necessary for adoption.
3. **Hybridity is Key:** The most successful projects use AI for the "heavy lifting" (initial transcription) but rely on human elders for the "fine tuning" (cultural validation).

5. Conclusion

The role of AI in preserving endangered Indigenous oral literatures is a paradox of power. It possesses the computational capacity to rescue languages from the brink of silence, archiving in a decade what would previously have taken a century. Yet, without a framework of Indigenous Data Sovereignty, AI risks becoming a tool of assimilationflattening diverse dialects into a standardized, "machine-readable" output.

The future of this field lies not in Silicon Valley, but in the "digital marae" and the reservation computer lab. The technology must remain a *servant* to the oral tradition, not its master. As we move forward, the metric of success for AI in this domain should not be how many languages it can translate, but how many speakers it creates

Works Cited

Bird, Steven. "Decolonising Speech and Language Technology."*Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 3504–3519.

Hermann, Isabella. "Artificial Intelligence in Fiction: Between Narratives and Metaphors."*AI & Society*, vol. 38, no. 1, 2023, pp. 319–329.

Jones, Peter, and Keoni Mahelona. *Papa Reo: Kaitiakitanga of Language Data*. Te Hiku Media, 2022.

Meta AI. "No Language Left Behind: Scaling Human-Centered Machine Translation."*Meta Research*, 2022, research.facebook.com/publications/no-language-left-behind/.

Smith, Linda Tuhiwai. *Decolonizing Methodologies: Research and Indigenous Peoples*. 3rd ed., Zed Books, 2021.

UNESCO. "Global Action Plan of the International Decade of Indigenous Languages (IDIL 2022-2032)."*UNESCO*, 2021, en.unesco.org/idil2022-2032/global-action-plan.

United Nations. *Declaration on the Rights of Indigenous Peoples*. UN Department of Economic and Social Affairs, 2007.