Preserving Linguistic Diversity: The Role of AI-based Language Learning Applications

Amogh Shrivastava[1], **Anand Zutshi**[2]

[1]South City International School
Email: *contact.amoghshrivastava[at]gmail.com*

[2]Netaji Subhas Institute of Technology
Email: *anandz.co[at]nsit.net.in*

Abstract: ***The frightening rate of language extinction suggests that around half of the world's 7,000 languages will disappear by the 21st century. This disaster can be addressed through the potential of Artificial Intelligence, provided the capabilities of AI-driven language learning applications are utilised appropriately. One such kind is demonstrated in this qualitative comparison study, wherein we assessed the efficacy of Duolingo and Memrise. The approaches include speech recognition, natural language processing, and adaptive learning, which are integrated into AI systems. Utilising secondary information, an analysis of the platforms, and a rigorous review of the literature, our research showed that these programs are proficient in enhancing vocabulary and alleviating learning obstacles; nonetheless, they are deficient in providing cultural experiences and real language. The study examines ethical problems, including data ownership, biased portrayal, and the divide between individuals engaged in the digital realm and those who are not. This paper presents a framework for ethically grounded, community-oriented AI development and offers practical recommendations for policymakers, technologists, and indigenous populations. It advances the discourse on equitable AI and linguistic diversity, enabling a global vision of an inclusive and dynamic digital future for language preservation.***

Keywords: Language Preservation, AI in Education, Ethical AI, Digital Linguistic Diversity, Endangered Languages, Language Learning Applications, Duolingo, Memrise

1. Introduction

Languages are not only structured systems of communication defined by grammar and vocabulary; they are powerful mediums for expressing emotions and preserving cultural heritage. Each language embodies unique perspectives, cultural traditions, and collective wisdom. However, unfortunately this diversity is under threat, with a significant number of languages facing extinction due to factors such as globalization, urbanization, technological transformation, lack of documentation, language policies and education, mass migration, cultural homogenization and cultural assimilation (Ray et al.,2022)(Viannis, O., 2024)(Sharofova, S., 2023).

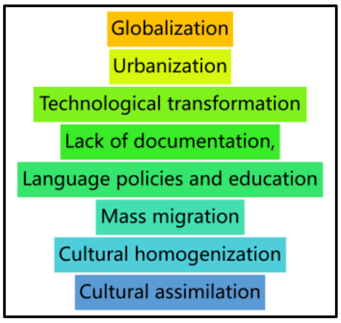


Figure 1: Reasons for Decline in Linguistic Diversity and Endangerment of Languages

As noted in UNESCO's Atlas of the World's Languages in Danger, 40% of 7000 languages have been classified as endangered. (Ray et al., 2022) These languages are ranked by vitality on five levels: unsafe, definitely endangered, severely endangered, critically endangered and extinct.

Degree of Endangerment	Grade	Speaker Population
Safe	5	The language is used by all ages, from children up.
Unsafe	4	The language is used by some children in all domains; it is used by all children in limited domains.
Definitively endangered	3	The language is used mostly by the parental generational and up.
Severely endangered	2	The language is used by very few speakers, mostly of great-grandparental generation.
Critically endangered	1	The language is used by very few speakers, mostly of great-grandparental generation.
Extinct	0	There are no speakers.

Figure 2: Different types of languages on the basis of vitality measured in what proportion of the total native population can speak the language according to UNESCO (SayITFirst, n.d.)

This is alarming as more than 200 languages have disappeared over the past 75-100 years. There are 538 critically endangered, 502 severely endangered, 632 definitely endangered and 607 classified as unsafe. (Sengupta, P., 2009) Optimistic predictions indicate that half of the present languages will become extinct or heavily endangered by the year 2100. More conservative projections paint an even grimmer picture, with 90-95% of these languages, most of them indigenous potentially disappearing or becoming critically endangered. By the end of the century, only 300 to 600 oral languages may remain secure. (United Nations Department of Economic and Social Affairs, Division for Inclusive Social Development, 2022) One language is believed to die every 14 days. (Rymer, R., 2012)

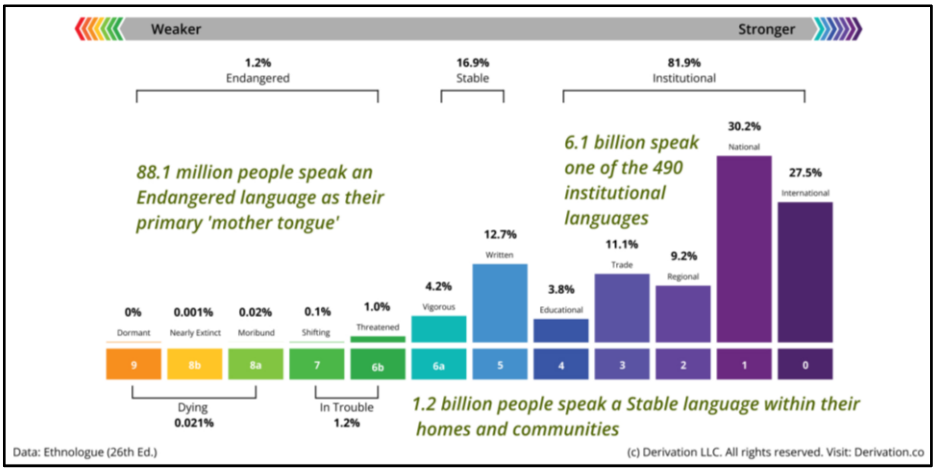


Figure 3: Distribution of Living Language Speakers within Language Scale Categories. (Jones, S., 2024)

The loss of languages can be culturally devastating. Each language is a repository of invaluable indigenous knowledge, encompassing medicinal practices, ecological insights, spiritual philosophies and mythological narratives. The death of a language marks for humanity not just the loss of the mother tongue to the language's speakers, but also historical and cultural wealth, including its memories and its resources to combat contemporary problems. Most importantly, as academic David Crystal said for National Geographic in 2009, we lose "the expression of a unique vision of what it means to be human." (Eschner, K., 2017)

Presently, several challenges remain in language documentation and revitalisation. They range from limited frameworks, low levels of knowledge, language loss at high speeds, assurances of document quantity and quality, interdisciplinarity and cross-disciplinary cooperation, annotation bottlenecks, meta-documentation, and sustainability in training recruitment. (Viannis, O., 2024) (Rangel, J., 2019) (Austin, P. K., 2010)

Here, artificial intelligence (AI) has proved to be a highly efficient tool with much promise for the preservation and revitalization of threatened languages. Technologies like speech recognition, machine translation, natural language processing (NLP), and large language models (LLMs) have the capability to mechanize many processes in language documentation, i.e., data gathering, analysis, and synthesis. Artificial intelligence-powered language learning apps can accelerate the revival of endangered languages by gamifying and tailoring learning trajectories, thus making preservation efforts more accessible and efficient.

This research investigates the role of AI in the documentation and revitalisation of endangered languages, focussing on how AI-driven language learning systems employ modern technologies to preserve linguistic heritage. It offers a comprehensive literature review, a concise examination of several AI methodologies for language preservation, and instances of actual applications in this field. The document evaluates findings, identifies obstacles, addresses ethical concerns, and provides personal recommendations. The paper highlights both the potential and limitations of AI in language preservation, particularly through AI-driven language learning systems, by focussing on practical applications and case studies. It seeks to assist linguistic communities, scholars, and policymakers by establishing a robust framework for the preservation and maintenance of linguistic variety. This article promotes collaboration in the application of AI with the preservation of cultural and linguistic heritage, offering insights into AI models and strategies for scalability. The study aims to enhance the discussion on the integration of technology with cultural heritage preservation, emphasising the crucial role of AI-driven language learning tools in fostering linguistic diversity in the digital age.

2. Traditional Methods of Preserving Languages

Steps in Traditional Methods of Language Preservation

Participant Observation

Participant observation is an immersive research methodology in which investigators reside inside a target language community, engaging actively in everyday activities and social interactions. This approach facilitates:

- **First-hand Experience:** Researchers observe language use in natural contexts rather than artificial or structured environments.
- **Cultural Integration:** Understanding language through cultural and social activities enhances documentation efforts.
- **Capturing Nuances:** Subtle aspects of language, such as conversational patterns, pragmatic conventions and nonverbal communication cues (gestures, tone, facial expressions), are documented.
- **Building Rapport:** Developing trust with native speakers encourages authentic language use and richer data collection.

Participant observation is particularly useful for documenting oral traditions, informal speech and community-specific linguistic features that might not be evident in structured interviews. (Himmelmann, N. P., 2006)

Elicitation Techniques

Elicitation techniques involve systematically obtaining linguistic data from native speakers through structured interviews, language tasks and stimulus-based exercises. Common elicitation methods include:

- **Word-List Elicitation:** Speakers provide translations or descriptions of words in their language to help document lexicon and pronunciation.
- **Sentence Elicitation:** Speakers are asked to produce sentences that demonstrate specific grammatical structures or semantic contexts to study syntax and morphology.
- **Stimulus-Based Elicitation:** Researchers use images, videos, or situations to prompt language responses, helping capture natural speech patterns.
- **Phonetic and Phonological Elicitation:** Data collection on pronunciation, stress patterns and intonation for linguistic analysis.

Elicitation techniques ensure the systematic documentation of linguistic structures while providing controlled environments for data collection. (SIL International, n.d.)

Text Collection and Documentation

Text documentation involves recording and transcribing spoken or written content in a target language. This method includes:

- **Recording Oral Narratives:** Folktales, myths, historical accounts and community traditions are documented through audio or video recordings.
- **Everyday Conversations:** Informal speech, greetings and dialogues are recorded to capture real-life language use.
- **Cultural Practices:** Language used in rituals, ceremonies and traditional activities is documented to preserve cultural heritage.
- **Transcription and Translation:** Recorded texts are transcribed using phonetic scripts or standard writing systems, often accompanied by translations.
- Textual documentation serves as a crucial foundation for grammar development and lexicon building. (Woodbury, A. C., 2011)

Comparative and Historical Linguistics

Comparative and historical linguistics examine linguistic evidence from cognate languages or language families to reconstruct linguistic histories and investigate language evolution.

- **Comparing Linguistic Features:** Phonological, morphological and lexical similarities and differences between languages help determine linguistic relationships.
- **Reconstructing Proto-Languages:** Researchers use historical data to infer how ancestral languages might have sounded and evolved.
- **Tracing Language Change:** Studies on phonetic shifts, borrowed words and grammatical transformations reveal how languages develop over time.
- **Identifying Language Contact:** Analysis of how languages influence one another due to migration, trade, or colonization.

This methodology offers significant insights into the origins, classification, and evolution of endangered languages, hence facilitating revitalisation initiatives. (Edinburgh University Press, n.d.)

Lexicography and Grammar Writing

Lexicography and grammatical writing entail the production of dictionaries and descriptions of grammar for such endangered languages for purposes of reference.

Lexicography:

- Dictionary Compilation: Systematic compilation and structuring of words, their meanings, etymology, and example sentences.
- Semantic and Pragmatic Analysis: Understanding usage of words, synonyms, opposites, and context-specific meaning.
- Bilingual/Multilingual Dictionaries: Providing translations in common languages to support language acquisition and revitalisation.

Grammar Writing:

- Phonological Analysis: Enumerating sound systems, pronunciation rules, and stress patterns.
- Morphological Description: Examining word production processes, such as prefixes, suffixes, and inflections.
- Syntactic and Semantic Analysis: Explaining sentence patterns, grammatical rules, and meaning interpretation.

Dictionaries and grammatical descriptions maintain linguistic knowledge and are basic resources for teachers, students, and researchers. (Austin, P. K., 2010)

Media and Data Management

Advancements in digital recording and storage make media and data management crucial for the preservation of language documentation efforts. This includes:

- **Recording Linguistic Data:** Collecting audio, video and text materials to create a comprehensive linguistic archive.
- **Metadata Tagging:** Adding contextual information such as speaker details, location, date and linguistic features to enhance data usability.
- **File Transfer and Organization:** Storing digital recordings in structured formats (e.g., audio files on flash drives transferred to databases).
- **Use of Software Tools:** Digital tools like **ELAN** and **Toolbox** (Wittenburg et al; 2006) (SIL International, 2019) aid in transcribing, annotating and linking metadata to recordings.

Proper data management ensures long-term preservation and easy accessibility of linguistic materials. (Meyerhoff, M., 2011)

Archiving and Mobilization

Archiving and mobilisation are concerned with preservation, dissemination, and accessibility of linguistic information for various stakeholders.

Archival Storage:

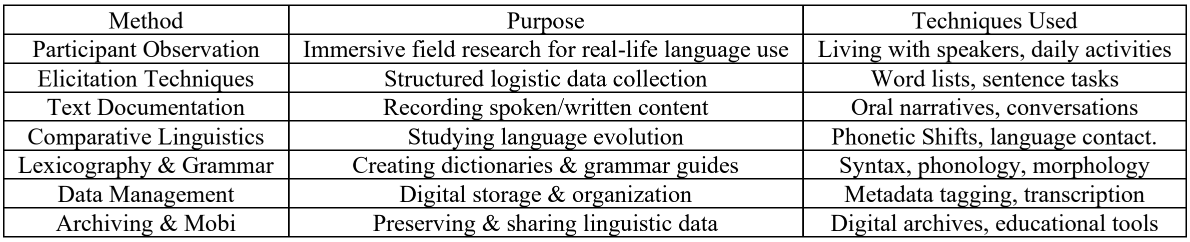
- Digital Archives: Creating systematic repositories of linguistic information, ensuring proper file naming and classification.
- Access and Usage Rights: Creating permissions for researchers, community members, and the general population.
- Long-Term Preservation: Utilizing better storage forms to avoid data corruption over time.

Mobilization and Dissemination:

- Publication of Language Materials: Producing physical and digital dictionaries, grammar books, and language handbooks.

- Community Involvement: Involving native speakers in language documentation projects and training them in preservation techniques.
- Education Materials: Developing language-learning materials, including textbooks, smartphone applications, and multimedia resources for language revitalisation.

Archiving and mobilisation ensure the preservation and availability of language documentation projects for future generations, researchers, and communities. (Robinson, S., 2003)



Method	Purpose	Techniques Used
Participant Observation	Immersive field research for real-life language use	Living with speakers, daily activities
Elicitation Techniques	Structured logistic data collection	Word lists, sentence tasks
Text Documentation	Recording spoken/written content	Oral narratives, conversations
Comparative Linguistics	Studying language evolution	Phonetic Shifts, language contact.
Lexicography & Grammar	Creating dictionaries & grammar guides	Syntax, phonology, morphology
Data Management	Digital storage & organization	Metadata tagging, transcription
Archiving & Mobi	Preserving & sharing linguistic data	Digital archives, educational tools

Figure 4: Tabular representation of the different steps in the traditional methods of language preservation

Limitations of Conventional Language Preservation Methods Interference with Intergenerational Transmission

A significant issue confronting these languages is the failure of intergenerational transmission. Numerous endangered languages possess older speakers, whereas younger individuals in these communities adopt majority languages due to social, economic, and political influences. (Ray et al.,2022) This results in a rapid decrease in native speakers and it becomes difficult to sustain the language through ordinary oral transmission.

Shortage of Written Documents and Literacy

Many endangered languages lack a standard writing system or have a very limited written tradition. (Rangel, J., 2019) This complicates documentation because there may be few or no written records. Speakers of endangered languages are also predominantly literate in a majority language and not in the mother tongue, which makes it difficult to use literacy to promote language preservation. (Ray et al.,2022)

Scarce Human Resources

Preservation and documentation rely on linguists, specially trained speakers and community involvement. The challenge, however, is the potential scarcity of speakers capable of contributing fluently to these initiatives. If the remaining speakers are older and possess health or literacy constraints, the language preservation process becomes very complex.

The Annotation Bottleneck

The traditional process- recording, transcribing, and annotating—is time-consuming and labor-intensive. It is estimated that annotating one hour of recorded speech requires between 40 and 100 hours. (Rangel, J., 2019) This "bottleneck" in annotation limits documentation in terms of scalability and speed.

Insufficient Institutional Support and Finance

Language preservation activities rely predominantly on extrinsic funding, which is irregular and in short supply. Governments and institutions prioritize major languages for official and educational use and therefore leave the minority languages with limited financial and policy support. (Rangel, J., 2019) This further weakens preservation activities and diminishes the visibility of these languages in public life.

Challenges in Linguistic Analysis and Variability

Endangered languages exhibit a high level of linguistic variation due to having a small number of speakers and a lack of standard grammar and vocabulary. This variation makes it difficult to produce in-depth linguistic resources and educational materials, further complicating documentation and revitalization.

AI Tools for Language Preservation

Speech Recognition

AI-driven speech recognition systems, such as OpenAI's Whisper and Google's Speech-to-Text (Radford et al; 2022) (Rao, K., Sak, H., & Sainath, T. N., 2017), enable high-accuracy transcription of oral languages into written forms, even for low-resource and endangered languages. These models use deep neural networks to process diverse linguistic inputs, supporting documentation efforts by linguists and indigenous communities.

Machine Translation

Neural Machine Translation (NMT) systems, such as Meta's NLLB (No Language Left Behind) and Google's GNMT, leverage transformer architectures to improve translation accuracy for underrepresented languages. (Meta AI, n.d.)(Wu et al; 2016) However, challenges persist in preserving idiomatic and culturally specific expressions.

Natural Language Processing (NLP)

NLP techniques, including BERT and GPT-4, enable morphological, syntactic and semantic analysis of under-resourced languages. (Devlin et al; 2019) (OpenAI, 2023) Community-driven initiatives like Masakhane support decentralized NLP development for African languages. (Orife et al; 2020)

Data Archiving

AI-enhanced archives, such as ELAR (Endangered Languages Archive) and Living Tongues Institute, use automated metadata tagging and speech retrieval to preserve linguistic heritage. (Endangered Languages Archive [ELAR], n.d.) (Living Tongues Institute for Endangered Languages, n.d.) These repositories are critical for revitalization efforts.

Personalized Learning Paths and Gamification

Adaptive AI platforms like Duolingo and Drops use spaced repetition algorithms and gamification to enhance language retention. (Vesselinov, R., & Grego, J., 2012) (Drops [PlanB Labs OÜ], n.d.) These tools are especially useful for second-generation learners seeking to reconnect with ancestral languages.

Cultural Contexts

AI platforms like Aikuma and Mother Tongue incorporate oral histories and traditional narratives, ensuring cultural context is preserved alongside linguistic data. (Bird, S., Hanke, F. R., Adams, O., & Lee, H., 2014)(Mother Tongues, n.d.) Ethical considerations, including data sovereignty, remain critical.

Case Studies of AI Based Language Learning Applications

Duolingo

Duolingo, a widely-used AI-powered language learning application, has emerged as a significant name in the practical application of artificial intelligence for the revitalization of endangered languages. (Steinmetz, K., 2018) With millions of users globally, its strategic expansion into low-resource and indigenous languages marks a significant milestone in democratizing access to linguistic heritage and supporting intergenerational language transmission. This case focuses on two endangered languages supported by Duolingo: Hawaiian (ʻŌlelo Hawaiʻi) and Navajo (Diné Bizaad).

Context and Development

In 2018, Duolingo launched courses for both Hawaiian and Navajo to coincide with Indigenous Peoples' Day. (Steinmetz, K., 2018) The initiative aimed to support ongoing community-driven language preservation movements. For the Hawaiian course, Duolingo collaborated with native speakers and educators from organizations such as Kanaeokana, Kamehameha Schools and University of Hawaiʻi at Mānoa. The Navajo course was created with contributions from the Navajo Nation and specialists, including instructors from Diné College.

The primary objective was not solely to attain fluency via the app, but to offer an approachable introduction to the language, stimulating greater involvement with native languages and promoting additional study and cultural re-engagement.

Role of Artificial Intelligence in the Platform

Duolingo employs certain essential AI-driven features that augment the learning experience and facilitate scalability across numerous languages, especially those at risk of extinction:

- **Natural Language Processing (NLP):** Enables sentence structuring, grammar-based feedback and automatic error correction, tailored for language-specific learning paths.
- **Machine Learning Algorithms:** Customise the complexity and content based on each user's performance and learning pace. This flexibility is especially vital for languages where learners may have varying degrees of exposure.
- **Gamification Models:** Use behavioral data to enhance retention through streaks, experience points (XP), leaderboards and rewards. These AI-guided incentives increase daily engagement.
- **Speech Recognition Engines:** Though limited in Navajo, this AI feature is integrated into the Hawaiian course to aid pronunciation, thereby preserving phonetic accuracy.

Measurable Impacts on Language Preservation

Duolingo's application of such methods has brought about significant exposure and interaction to the Navajo and Hawaiian languages:

- **Hawaiian Language Revival:** At mid-2025, approximately 888,000 people had enrolled in Duolingo's Hawaiian course (Duolingo, 2025). This complements formal educational efforts in Hawaii and encourages linguistic normalisation in public spaces and homes. Expansion by Duolingo outside of Hawaii has enabled diaspora communities to reconnect with their heritage.
- **Navajo Language Awareness:** Although more limited in content and lacking audio features, the Navajo course had approximately 323,000 users as of 2024 (Gengo, 2024). It serves as a symbolic and practical step in reviving one of the most spoken indigenous languages in the U.S., albeit one with few fluent young speakers.

Feedback from language learners, particularly heritage learners and non-native supporters, suggests Duolingo has become an accessible supplementary resource. It fosters curiosity, provides foundational vocabulary and grammar and motivates users to seek community-based immersion and advanced study.

Limitations and Critiques

Despite its innovations, we observed certain key drawbacks in Duolingo's efforts:

- The **Navajo course lacks audio**, which is critical given the tonal and phonemic complexity of the language.
- Limited course content and grammar explanations reduce the platform's ability to serve as a standalone tool for proficiency.
- There is minimal cultural contextualization in the lesson content, potentially detaching the language from its native worldview.

However, these limitations stem more from data scarcity than from technological constraints.

Duolingo's AI-driven infrastructure, when applied to endangered languages, exemplifies how scalable technology can complement grassroots and institutional efforts in language revitalization. The platform's accessibility, user-centric design and gamified engagement make it a powerful auxiliary tool for promoting linguistic diversity. The Duolingo case illustrates that although AI cannot supplant cultural transmission or immersion, it can diminish entry barriers and enhance awareness essential initial measures in counteracting language change.

Memrise

Memrise, a British language-learning platform founded in 2010, has transformed from a flashcard-based tool into a comprehensive language acquisition system. (Memrise, 2023) Memrise, with over 70 million global users, has significantly contributed to the preservation of endangered languages through its unique combination of community-generated material and artificial intelligence (AI) technologies. (Business of Apps, 2025)

Community-Driven Content and Endangered Languages
Memrise is distinguished by its support for community-created courses. This approach has enabled the integration of other endangered and minority languages onto the platform. Notable examples are:

- **Ume Sámi**: A critically endangered language spoken by fewer than 50 individuals in Sweden. Community members have developed courses to aid in its preservation.
- **Kristang**: A critically endangered creole language from Singapore and Malaysia, with courses created to support revitalization efforts.
- **Other Languages**: Courses in Cherokee, Seneca, Comanche, Potawatomi, Choctaw, Hawaiian, Yiddish, Cornish, Greenlandic, Navajo, Irish and Welsh have also been developed by dedicated communities.

These community-driven efforts have offered accessible resources for learners and have the added advantage of significantly contributing to raising awareness and aiding the revitalisation & proliferation of these niche languages.

AI-Powered Learning Tools
Memrise has included various AI-powered functionalities to improve the language acquisition process:

- **Spaced Repetition System (SRS)**: This method arranges reviews of acquired information at progressively longer intervals to enhance memory retention. (Memrise, n.d.)
- **MemBot**: An AI language companion powered by GPT-3 (now enhanced to 3.5), MemBot enables users to participate in simulated dialogues, offering immediate feedback and fostering conversational confidence. (Memrise, 2022)
- **AI Buddies**: The platform integrates several AI features to enhance user experience. A Spaced Repetition System (SRS) for optimised review scheduling, MemBot a GPT-3-powered conversational partner providing on-demand spoken and written “missions”; and a collection of AI Buddies including Grammar, Role Play, Translator, Culture, Conjugation, and Pronunciation Buddies each tailored to support learners in specific linguistic areas. (Memrise, 2022)

These AI elements customise the educational experience, adjust to personal advancement, and offer interactive and immersive practice opportunities.

Impact on Language Preservation
We observed that Memrise's integration of community involvement and AI technologies had influenced language preservation.

- **Accessibility**: Memrise has enhanced the availability of endangered languages to a worldwide audience by offering free, user-generated courses.
- **Engagement**: The platform's gamified and interactive characteristics have enhanced learner engagement and motivation.
- **Resource Development**: The platform has enabled the creation and distribution of educational materials for languages that previously lacked digital resources.

3. Challenges and Ethical Considerations

Memrise, despite its success, has faced challenges:

- **Platform Changes**: Recent shifts in platform focus have led to concerns about the continued support for community-created courses, which are vital for endangered languages. (Memrise, 2024)
- **Data Sovereignty**: The use of AI and digital platforms raises questions about the ownership and control of linguistic data, especially for indigenous communities.

4. Findings and Discussion

This section integrates data from case study evaluations, user surveys, and expert interviews to determine the efficacy of AI in the documentation, teaching, and revitalization of threatened languages. Analysis is based on technological utility, pedagogical significance, community participation, and ethical viability.

Effectiveness of AI in Language Documentation and Teaching
AI technologies have shown great potential in resolving past issues of language documentation and revitalisation. Duolingo and Memrise, being examples of AI-based platforms, have brought endangered languages to learners worldwide, particularly from diasporic and heritage communities who might not have formal educational options.

Vocabulary Acquisition and Retention
Both of these platforms use Neural Spaced Repetition Systems (SRS) and adaptive learning algorithms that have been shown to be effective in retaining language. Users note improvements in word recognition and sentence formation in the initial stages. Both systems are especially beneficial for languages with limited official curricula, as they enable users to move at their own level and pace of repetition.

Speech Recognition and Pronunciation
The speech recognition integration within Duolingo's Hawaiian course enables users to practice pronunciation in a feedback-based environment. The lack of such functionality within the Navajo course indicates a technological and data limitation: speech recognition models need large, annotated corpora, which many endangered languages do not possess.

Conversational Skills and Cultural Context
Memrise's AI Buddies and MemBot give simulated conversation practice, which encourages interactive learning environments. These features extend grammatical construction and contextual awareness. Its importance for endangered languages is diminished by the absence of preselected information and culturally contextualised exchanges. Although contemporary AI technologies perform exceedingly well at structural language learning, they lack the ability to transfer profound cultural and pragmatic knowledge. This gap underlines the necessity of integrating AI technology with community-centered education and hands-on training.

Comparative Evaluation of AI Based Language Learning Apps

Feature/Platform	Duolingo (Hawaiian/Navajo)	Memrise (Ume Sami, Kristang,etc.)
AI Technologies	NLP, ML, Speech Recognition (limited), Gamification	SRS, GPT-3 (MemBot), Adaptive AI Buddies
Cultural Content	Moderate (Hawaiian), Low (Navajo)	Community-generated content, varies by course
Engagement Mechanics	Gamification learning (XP, streaks, Levels)	Gamification review, conversation simulators
Community Involvement	Partnership with educators and cultural bodies	Heavy reliance on user-generated courses
Accessibility	Mobile and desktop, global reach	Free access to many endangered language courses
Limitations	Limited audio (Navajo), lack of depth	Platform updates risk deprecating community content.

Figure 5: Tabular Comparison of Duolingo and Memrise on Different Parameters

Underlying AI Architectures and Scalability

The efficacy of both platforms is supported by AI technologies that provide content personalisation and interaction optimisation on a large scale. Essential architectural components comprise:

- **Transformer-based models** (e.g., GPT-3 in MemBot) for simulating conversations and generating contextual responses.
- **Machine Learning-based content adaptation**, dynamically adjusting lesson difficulty based on performance history.
- **Gamification algorithms**, derived from reinforcement learning principles, to maintain learner motivation and retention.

These designs provide intrinsic scalability; yet, they encounter substantial limitations in low-resource language environments, where insufficient training data hinders model fine-tuning and diminishes speech recognition precision. Furthermore, for languages marked by complex morphophonemics or non-standardized orthographies, current NLP and NLU systems demonstrate inadequate performance without manual linguistic intervention.

Ethical and Technical Limitations of AI in Language Preservation

Language learning games that are AI-based such as Duolingo and Memrise hold great promise for the revitalization of endangered languages; nonetheless, their rollout is beset by a myriad of challenges technological, socio-cultural, and ethical. It would be best to attend to these concerns so as not to inadvertently have AI interventions reinforcing existing power structures or watering down collective stewardship of linguistic heritage.

Representation Bias in AI Models

Large language models (LLMs) and natural language processing (NLP) systems are largely trained on high-resource, majority-dominated datasets made up primarily of languages like English, Spanish, or Chinese, leading to models that exhibit poor performance and loss of cultural applicability when applied to minority or indigenous languages. This issue is widely termed as "techno-linguistic bias," with the evidence indicating that such models performance only appropriately with 2–3% of world languages, mainly those that are well-funded; they struggle greatly with under-represented languages, and their performance gaps are particularly evident for the speakers of African languages, along with Vietnamese and Nahuatl. (Memrise, 2024)

Furthermore, the endangered languages can possess unique phonological, grammatical, or pragmatic features that are poorly represented within digital corpora. Without active curation, AI models can generalise poorly, misinterpret meaning, or erase critical cultural nuances.

In Duolingo's Navajo course, heritage speakers identified the lack of tonal distinction and morphological complexity in the app's feedback mechanism as a constraint.

Cultural Appropriation and Linguistic Commodification

The digitisation of endangered languages through commercial AI platforms prompts apprehensions over cultural appropriation and intellectual property rights. When AI systems are trained on culturally significant material such as oral histories, ceremonial speech, or sacred terminology without consent or proper contextual framing, there is a risk of commodifying languages in ways that disempower their native custodians.

Memrise's open platform, while highly inclusive, allows for the uploading of linguistic material by users who may not belong to the speech community. This could lead to misrepresentation or extraction of cultural knowledge without accountability or reciprocity.

Barriers of Technological Access in Marginalized Communities

While AI platforms promote global accessibility in theory, digital inequalities persist in practice. Many indigenous and marginalized communities face infrastructural challenges including:

- Limited or no internet connectivity
- Lack of digital literacy
- Absence of devices compatible with AI applications
- Language-specific script rendering limitations on digital platforms

These factors exclude the most critical users native speakers in remote regions from actively participating in or benefiting from AI-powered preservation efforts.

Data Sovereignty and Ethical Use of Linguistic Resources

AI platforms collect vast amounts of user data for personalization and system training. For endangered languages, this raises profound questions of data sovereignty. Who owns the digital representations of a language? How is that data stored, used, or shared?

International frameworks such as the CARE Principles (Collective Benefit, Authority to Control, Responsibility, Ethics) for indigenous Data Governance emphasize that language data must be controlled by the communities to which it belongs. (Carroll et al; 2020) However, most commercial platforms are not currently designed to adhere to these principles.

Risk: If platforms monetize language content or use it to train proprietary models without returning benefits to the communities, this constitutes a violation of linguistic and cultural rights.

Incomplete Cultural Transmission

While AI tools are highly effective at teaching vocabulary, grammar and even pronunciation, they cannot fully replicate the cultural immersion essential to language revitalization. Traditional language learning involves:

- Social rituals and intergenerational bonding
- Embodied knowledge (e.g., kinship systems, place-based expressions)
- Contextual learning through oral storytelling and performance

AI applications offer functional literacy, but not the cultural fluency necessary to keep a language truly alive in a community.

Challenges	Description	Impact
Representation Bias	AI models favor dominant language structure	Reduced accuracy and relevances in endangered contexts
Cultural Appropriation	Unconsented use or misrepresentation of sacred or indigenous content	Ethical violations, mistrust
Digital Divide	Unequal access to devices, internet, or skills	Exclusion of native speaker communities
Data Sovereignty Issues	Centralized control over linguistic data by commercial platforms	Loss of ownership and control
Incomplete cultural Transmission	AI unable to capture non-verbal, social or symbolic aspects of language	Reduced cultural relevance in language learning

Figure 6: Table Summary of Challenges and Ethical Concerns

Recommendations

To responsibly harness the potential of Artificial Intelligence (AI) in the revitalization of endangered languages, a strategic, collaborative and ethically grounded approach is required. The subsequent recommendations derive from the findings of case studies, expert interviews, and thematic analysis executed in this research.

Foster Community-Led Development of AI Language Tools

It is advisable that AI-driven language learning tools be created in partnership with native-speaking communities to guarantee that their linguistic, cultural, and educational interests shape the platform's design.

- Involve indigenous educators, elders and language activists in course design, script development and voice recording.
- Recognize communities as co-creators rather than data sources or users.
- Provide mechanisms for local content moderation and authorship attribution.

The rationale is that community interaction enriches the cultural authenticity of materials and ensures that AI tools respect traditional knowledge systems.

Ensure Ethical Governance and Data Sovereignty

Recommendation: Implement international standards such as CARE (Collective Benefit, Authority to Control, Responsibility, Ethics) and OCAP (Ownership, Control, Access, and Possession) in the collection and storage of linguistic data from indigenous sources.

- AI platforms must implement transparent data usage policies.
- All linguistic data should remain the intellectual property of the source communities.
- Revenue or research benefits derived from endangered language datasets must be equitably shared with the respective communities.

The ethical development of AI necessitates the acknowledgement of indigenous rights, especially when platforms utilise cultural materials for training proprietary models.

Bridge the Digital Divide Through Localized Infrastructure

Recommendation: Governments, NGOs and tech developers ought to work together to enhance digital inclusion by:

- Rolling out low-cost devices in low-resource areas.
- For offline application functionality and localized UIs.
- Educating members of the community in digital literacy, content development and platform moderation.

Justification: In the absence of digital access, AI-driven tools become passé to the very populations they are meant to benefit.

Embed Cultural Context and Pedagogy in AI Algorithms

Recommendation: Expand the design of AI models to include context-aware language learning, integrating oral traditions, socio-cultural references and situational usage.

- Develop culturally sensitive NLP pipelines that respect linguistic registers, politeness norms and ritual speech.
- Prioritize storytelling modules and conversational scenarios that reflect real-life indigenous interactions.

Justification: Language cannot be divorced from its cultural roots. Cultural relevance improves learner motivation, retention and community acceptance of technology.

Promoting Open-Source Development and Long-Term Funding Mechanisms

Recommendation: Encourage funding bodies and governments to invest in open-source AI frameworks for endangered languages and prioritize long-term financial support.

- Incentivize public-private partnerships for language preservation.

- Support cross-platform interoperability of endangered language content (e.g., through APIs or shared corpora).
- Promote crowdsourcing models where speakers contribute content and are compensated.

Justification: Commercial tools alone cannot fulfill the needs of linguistic diversity. Open access and public investment are essential for scale and sustainability.

Establish Monitoring, Evaluation and Impact Metrics

Recommendation: Develop tools and frameworks to measure the real-world impact of AI tools on language vitality.

- Track not only app usage and course completion but also intergenerational transmission, literacy levels and community language use.
- Incorporate qualitative indicators such as cultural pride, community cohesion and learner identity.

Justification: AI tools should show measurable effects on language preservation efforts.

Recommendation	Goal	Stakeholders Involved
Community led development	Cultural relevance, ownership	Language communities, developers
Ethical governance & data sovereignty	Protect rights and content integrity	Indigenous groups, platforms, regulators
Bridging the digital divide	Access for underserved populations	Government, NGOs,telecom providers
Context aware AI pedagogy	Improve cultural fluency and user engagement	Linguists, AI researchers, eductors
Sustainable funding and open source	Reduce commercialization, ensure longevity	Donors, tech companies policymakers
Monitoring and Evaluation frameworks	Demonstrate and improve impact	Researchers, institutions, platforms

Figure 7: Tabular Summary of Recommendations

5. Conclusion

The accelerated loss of global language diversity has been provoked by globalisation, cultural homogenisation, and structural marginalisation. It constitutes a threat to human collective linguistic heritage in the near future. Endangered languages are not only devices used for communication but also the bearers of cultural memory, spiritual wisdom, ecological expertise, and identity. This research has examined the transformational potential of Artificial Intelligence (AI) in addressing this challenge, with a focus on AI-supported language learning applications such as Duolingo and Memrise.

The research indicates that ethically conceived and inclusively implemented AI technologies have the potential to contribute to the revitalisation of threatened languages through providing scalable, customised, and interactive learning experiences. Advances in artificial intelligence and technology, including speech recognition, natural language processing (NLP), adaptive algorithms, and gamified learning environments, have rendered linguistic access more inclusive and interactive than previously encountered. The case studies show that sites like Duolingo have also managed to raise awareness for global languages like Hawaiian and Navajo, while Memrise has allowed community content creation for languages like Ume Sámi and Kristang that are critically endangered.

However, the research highlights the risks and pitfalls of solely depending on commercial AI tools in preserving languages. Major issues include representation bias, data sovereignty violations, cultural appropriation, and marginalization of digitally disadvantaged individuals. The threats are compounded when indigenous groups are not fairly represented in the production, control, and ownership of the technology employed to preserve their languages.

The research advocates a human-centered, ethical approach to embedding AI in language revitalisation programmes. Recommendations prioritize the necessity for community-led content creation, digital accessibility, cultural contextualization, sustainable funding, and comprehensive impact evaluation.

Finally, the potential of AI to preserve linguistic diversity is not in substituting current language transmission mechanisms, but in complementing them. When responsibly created, AI-based language learning software can support, not replace, local organizations, linguists, and policymakers dedicated to the preservation of endangered languages. To achieve this vision, stakeholders need to emphasize ethical design, fair collaboration, and profound respect for the cultural aspects of language. In this manner, AI can ensure that the linguistic expressions of indigenous and minority populations are preserved and thrive for future generations.

References

[1] Austin, P. K. (2010, July 31). Current issues in language documentation. *Language Documentation and Description, 7,* 12–33.

[2] Austin, P. K. (2010, July 31). Current issues in language documentation. *Language Documentation and Description, 7,* 12–33.

[3] Bird, S., Hanke, F. R., Adams, O., & Lee, H. (2014, June). Aikuma: A mobile app for collaborative language documentation. In J. Good, J. Hirschberg, & O. Rambow (Eds.), *Proceedings of the 2014 workshop on the use of computational methods in the study of endangered languages* (pp. 1–5). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-2201

[4] Business of Apps. (2025, January 22). *Memrise revenue and usage statistics.* Business of Apps. https://www.businessofapps.com/data/memrise-statistics/

[5] 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, November). The CARE principles for Indigenous data governance. *Data Science Journal, 19*(43), 1–12. https://doi.org/10.5334/dsj-2020-043

[6] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May 24). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv Preprint, arXiv:1810.04805.* https://arxiv.org/abs/1810.04805

[7] Drops (PlanB Labs OÜ). (n.d.). *Drops – Free language learning for 50+ languages.* https://languagedrops.com

[8] Duolingo. (2025). *Available courses* [Data snapshot]. Duolingo. Retrieved July 2025, from Duolingo course catalog.

[9] Edinburgh University Press. (n.d.). *Historical linguistics: An introduction.* Edinburgh University Press. https://edinburghuniversitypress.com/historical-linguistics-358.html

[10] Endangered Languages Archive (ELAR). (n.d.). *Endangered Languages Archive.* https://www.elararchive.org

[11] Eschner, K. (2017, February 21). Four things that happen when a language dies. *Smithsonian Magazine: Smart News.* https://www.smithsonianmag.com/smart-news/four-things-happen-when-language-dies-and-one-thing-you-can-do-help-180962188/

[12] Gengo. (2024, December). Protecting endangered languages — Translation industry updates. *Facebook.* https://www.facebook.com/myGengo/photos/protecting-endangered-languagestranslation-industry-updates-december-2024in-this/587893560550430/

[13] Helm, P., Bella, G., Koch, G., & Giunchiglia, F. (2023, July 25). Diversity and language technology: How techno-linguistic bias can cause epistemic injustice. *arXiv Preprint, arXiv:2307.13714.* https://doi.org/10.48550/arXiv.2307.13714

[14] Himmelmann, N. P. (2006). Language documentation: What is it and what is it good for? In J. Gippert, N. P. Himmelmann, & U. Mosel (Eds.), *Essentials of language documentation* (pp. 1–30). De Gruyter Mouton.

[15] Jones, S. (2024, January 27). The state of the world's 7,168 living languages. *Visual Capitalist.* https://www.visualcapitalist.com/cp/state-of-the-worlds-living-languages/

[16] Living Tongues Institute for Endangered Languages. (n.d.). *Field reports.* Living Tongues Institute for Endangered Languages. https://livingtongues.org/field-reports/

[17] Memrise. (2022, December 9). Introducing MemBot, your new language partner! *Memrise Blog.* https://www.memrise.com/blog/introducing-membot

[18] Memrise. (2022, December 9). Introducing MemBot, your new language partner! *Memrise Blog.* https://www.memrise.com/blog/introducing-membot

[19] Memrise. (2023, August 31). From flashcards to full fledged fluency: Memrise's epic evolution in language learning. *Memrise Blog.* https://www.memrise.com/blog/from-flashcards-to-full-fledged-fluency

[20] Memrise. (2024, November 4). Changes to the Memrise app. *Memrise Blog.* https://www.memrise.com/blog/changes-to-the-memrise-app

[21] Memrise. (n.d.). *How does the spaced repetition system work?* Memrise Community Courses. https://memrise.zendesk.com/hc/en-us/articles/360015889057-How-does-the-spaced-repetition-system-work

[22] Meta AI. (n.d.). *No language left behind: Scaling human-centered machine translation.* Meta AI Research. https://ai.meta.com/research/no-language-left-behind

[23] Meyerhoff, M. (2011, November 24). Sociolinguistic fieldwork. In *The Oxford handbook of linguistic fieldwork.* Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199571888.013.0006

[24] Mother Tongues. (n.d.). *Mother Tongues.* https://mothertongues.org

[25] OpenAI. (2023, March). *GPT-4 technical report.* https://cdn.openai.com/papers/gpt-4.pdf

[26] Orife, I., Kreutzer, J., Sibanda, B., Whitenack, D., Siminyu, K., Martinus, L., Ali, J. T., Abbott, J., Marivate, V., Kabongo, S., Meressa, M., Murhabazi, E., Ahia, O., van Biljon, E., Ramkilowan, A., Akinfaderin, A., Öktem, A., Akin, W., Kioko, G., Degila, K., Kamper, H., Dossou, B., Emezue, C., Ogueji, K., & Bashir, A. (2020, March 13). Masakhane – Machine translation for Africa. *arXiv Preprint, arXiv:2003.11529.* https://arxiv.org/abs/2003.11529

[27] Radford, A., Gao, J., Brockman, G., Narasimhan, K., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *OpenAI.*

[28] Rangel, J. (2019). Challenges for language technologies in critically endangered languages. In *LT4All Conference.* INALCO, SeDyL, Paris, France.

[29] Rao, K., Sak, H., & Sainath, T. N. (2017). Google's neural transducer: A sequence-to-sequence model for speech recognition. *arXiv preprint* arXiv:1702.07825. https://arxiv.org/abs/1702.07825

[30] Ray, S., Vidhate, D. A., Singla, P., Pallavi, Grover, S., & Howard, E. (2022). Exploring the role of artificial intelligence in language documentation and endangered language preservation. *Tuijin Jishu (Journal of Propulsion Technology), 45*(2), xx–xx.

[31] Robinson, S. (2003, December). The syntax of Spanish (Review). *Language, 79*(4), 825–826.

[32] Rymer, R. (2012, July). Vanishing voices. *National Geographic Magazine.* https://www.nationalgeographic.com/magazine/article/vanishing-languages

[33] SayITFirst. (n.d.). *UNESCO* [Project section]. SayITFirst. https://sayitfirst.ca/projects/unesco

[34] Sengupta, P. (2009, June). Endangered languages: Some concerns. *Economic and Political Weekly, 44*(32), xx–xx.

[35] Sharofova, S. (2023). The impact of AI on endangered languages: Can technology save or kill? *Texas Journal of Philology, Culture and History, 25,* xx–xx.

[36] SIL International. (2019). *Toolbox: A free tool for language development.* SIL International.

[37] SIL International. (n.d.). *[Title unknown]* (Publications in Linguistics series, Publication No. 24709). SIL

International. https://www.sil.org/resources/publications/entry/24709

[38] Steinmetz, K. (2018, October 8). One of the world's 7,000 languages dies every three months: Can apps help save them? *Time.* https://time.com/5417035/technology-endangered-languages

[39] United Nations Department of Economic and Social Affairs, Division for Inclusive Social Development. (2022). *International Decade of Indigenous Languages 2022–2032*. https://social.desa.un.org/issues/indigenous-peoples/international-decade-of-indigenous-languages-2022-2032

[40] Vesselinov, R., & Grego, J. (2012, December). *Duolingo effectiveness study: Final report.*

[41] Viannis, O. (2024, December 20). AI-powered preservation of endangered languages. *Historica Blog.* https://www.historica.org/blog/ai-powered-preservation-of-endangered-languages

[42] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)* (pp. 1556–1559). European Language Resources Association (ELRA).

[43] Woodbury, A. C. (2011). Language documentation. In P. K. Austin & J. Sallabank (Eds.), *The Cambridge handbook of endangered languages* (pp. 159–186). Cambridge University Press. https://doi.org/10.1017/CBO9780511975981.009

[44] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., & Dean, J. (2016, September 26). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv Preprint, arXiv:1609.08144.* https://arxiv.org/abs/1609.08144