REVIEW

AI-Powered Innovations for Documenting and Revitalizing African Languages

Jackton Otieno Midigo 🆔

Department of Languages, Linguistics and Literature, Gretsa University, Thika , Kiambu County P.O. Box 03-01000, Thika, Kenya

ABSTRACT

The documentation and revitalization of African languages are crucial for preserving the continent's linguistic and cultural heritage amid increasing threats of language endangerment. This study presents a systematic review of existing literature on artificial intelligence (AI)-driven approaches to language documentation and revitalization, adhering to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. Data were collected from twenty academic digital repositories and peer-reviewed journals specializing in computational linguistics, AI applications in language preservation, and African linguistics. Four major databases were specifically searched: Google Scholar, SpringerLink, ScienceDirect, and African Journals Online (AJOL). Peer-reviewed articles from established repositories were analyzed to explore key AI applications such as speech recognition, machine learning for transcription, neural machine translation, and digital archiving. The findings reveal that AI significantly enhances language preservation by enabling automated transcription, corpus development, and the creation of interactive learning tools. Nonetheless, challenges remain, including limited data availability, ethical concerns over language ownership, and technological accessibility in marginalized communities, which hinder widespread implementation. The study emphasizes the importance of interdisciplinary collaboration among linguists, AI developers, and local communities to ensure culturally sensitive and effective AI solutions. Ultimately, this review highlights the transformative potential of AI in supporting the sustainable revitalization of indigenous African languages and contributing to global linguistic resilience.

Keywords: African Languages; Artificial Intelligence; Language Documentation; Revitalization; Systematic Review; Linguistic Resilience

CORRESPONDING AUTHOR:

Jackton Otieno Midigo, Department of Languages, Linguistics and Literature, Rongo University, Rongo, Migori County P.O. Box 103-40404, Kenya; Email: jaktonmidigo@gmail.com; Tel.: +254722158739

1. Introduction

Africa is home to an estimated one-third of the world's languages, making it one of the most linguistically diverse continents on the planet [1]. According to Nairutia Kemei et al. [2], indigenous African languages serve not only as mediums of daily communication but also as repositories of cultural heritage, oral traditions, and indigenous knowledge systems. For instance, traditional songs, folktales, proverbs, and communal values serve as living archive of identity and continuity for the community. They play a vital role in expressing identity, fostering community cohesion, and transmitting traditional values and worldviews across generations [3]. However, this rich linguistic diversity is increasingly under threat due to the combined forces of globalization, urbanization, and the growing dominance of ex-colonial and global languages such as English, French, Portuguese, and Arabic in domains such as education, media, governance, and business [4].

As suggested by Adamou [5], the marginalization of indigenous languages has led to a gradual shift in language use patterns, especially among younger generations, resulting in language attrition and, in many cases, complete language loss. As suggested by Ogwudile [6], many African languages are now classified as endangered, with some on the verge of extinction. This trend not only poses a threat to Africa's linguistic heritage but also endangers the cultural, historical, and intellectual capital embedded in these languages [7]. In response, scholars, policymakers, and language activists have increasingly emphasized the urgent need for deliberate and effective documentation, revitalization, and promotion strategies aimed at preserving and restoring the vitality of endangered African languages.

In recent years, technological innovation, particularly in the field of Artificial Intelligence (AI), has emerged as a potential game-changer in the efforts to preserve linguistic diversity. Zhong et al. posit that AI technologies are being applied in various domains to solve complex global problems [8], including those in the field of linguistics and language preservation. According to Jermakowicz [9], AI is capable of transforming the way linguistic data is collected, analyzed, and disseminated. For instance, language models powered by AI can take spoken words and turn them into written form. This makes it much easier to capture the unique sounds and sentence structures that define endangered languages. Wang argues that advances in AI-driven tools such as automated transcription, corpus development, speech synthesis, and neural machine translation offer unprecedented opportunities for documenting and revitalizing endangered languages with greater efficiency and accuracy [10].

Rehm and Way further highlight that AI can facilitate sustainable linguistic resilience by enabling communities to create digital resources [11], develop language learning applications, and increase the accessibility and visibility of languages in digital and educational platforms. As a result, AI presents a promising frontier in linguistic research and advocacy, capable of complementing traditional language preservation methods with innovative, scalable solutions [12].

Given the growing urgency to preserve Africa's endangered languages and the emerging role of AI in this space, this study explores the integration of Artificial Intelligence (AI) as a technological media platform, specifically, a mediating tool that facilitates the processing, preservation, and dissemination of linguistic and cultural knowledge in the documentation and revitalization of indigenous African languages. By examining current practices, opportunities, and challenges, the research aims to contribute to the discourse on sustainable and technology-driven language preservation strategies tailored to the African context.

1.1. Statement of the Problem

The endangerment of indigenous African languages poses a serious threat to Africa's cultural and linguistic heritage. Despite efforts by linguists and policymakers, many of these languages remain under-documented and endangered, as traditional preservation methods lag behind rapid socio-technological change. Artificial intelligence (AI) offers new possibilities for language documentation and revitalization, but knowledge about its application remains limited. This study explores the role of AI in addressing the documentation and revitalization of African languages using AI by identifying research gaps, evaluating current applications, and proposing strategies to enhance AI-driven language preservation. It aims to support the development of innovative, inclusive, and sustainable approaches for safeguarding Africa's linguistic diversity.

1.2. Objectives

The main objectives of this study are:

1. To investigate the role of Artificial Intelligence (AI) in the revitalization of indigenous African languages.

2. To examine how Artificial Intelligence (AI) enhances systematic digital archives for indigenous languages.

3. To explore the key challenges impeding the effective implementation of AI-driven approaches in the documentation and preservation of African languages.

2. Literature Review

2.1. Language Revitalization

The preservation and revitalization of endangered languages have long been central to linguistic research, particularly as globalization continues to threaten the world's linguistic diversity. Foundational studies, such as those by Dennis [13], reinforce the importance of systematic documentation, capturing phonological, grammatical, and lexical structures not only for academic preservation but also for future revitalization initiatives. While this work provides essential methodological grounding, it remains anchored in traditional, human-led frameworks that may no longer suffice given the urgency and scale of language endangerment today. More critically, Dennis does not engage with the role of digital or AI-driven tools [13], missing a growing area of innovation that could enhance efficiency, scale, and community participation, especially in resource-constrained environments typical of many African linguistic contexts.

This omission is particularly significant when viewed in light of studies like those by Xuan and Yang [14], which highlight the potential of AI technologies, such as automatic speech recognition (ASR), machine translation, and corpus-building tools, to transform the documentation process. These technologies are especially useful in areas where linguistic expertise is scarce, and where fieldwork is constrained by geography, security, or funding. Yet, despite these advancements, the deployment of such tools in African contexts has often been met with uneven success. Some ASR systems, for instance, struggle with tonal languages or dialectal variability, both of which are common

in African linguistic landscapes. While AI may accelerate transcription, its effectiveness is heavily dependent on the quantity and quality of existing language data, resources often lacking for under-documented languages. The current study aims to bridge this gap by exploring how such tools can be adapted to the linguistic particularities of African communities, rather than assuming the transferability of models trained on major world languages.

In contrast to the technical focus of documentation, researchers emphasize the human dimension of language revitalization through community-led education programs, cultural events, and materials development [15,16]. These approaches ground language preservation in the lived realities of speakers, ensuring that revitalization is not abstracted into academic or institutional projects. However, while deeply rooted in community engagement, these studies remain largely analog in orientation. Their strategies often rely on face-to-face instruction or printed materials, assuming a level of resource availability and continuity that may not hold in dispersed or diasporic speaker populations. Additionally, such traditional revitalization frameworks risk alienating younger, tech-oriented generations who engage with language primarily through digital means, a generational shift that Tohit and Haque underscore as critical for any future-facing revitalization strategy [17].

This gap between traditional revitalization and digital engagement is further widened by the limited exploration of AI-enhanced language learning tools. Applications that incorporate gamification, speech synthesis, and immersive virtual environments could offer scalable and engaging ways to revive endangered languages, especially when tailored to local sociocultural contexts. Yet, as of now, few revitalization efforts within Africa have effectively integrated these technologies, either due to infrastructural challenges or a lack of culturally specific design. The current study seeks to evaluate how these tools can be made responsive to local realities, ensuring that technological intervention does not override or homogenize community values.

Adding a valuable cultural lens to the conversation, researchers frame language preservation as a fundamentally social and identity-driven act [18,19]. Their emphasis on intergenerational transmission, cultural continuity, and the embeddedness of language in everyday life pushes back

against the reduction of preservation to data collection or linguistic abstraction. However, while these works rightly center culture, they stop short of exploring how technology might be leveraged within these cultural systems rather than in opposition to them. As Alaimo and Kallinikos note [20], the production and circulation of knowledge today are increasingly shaped by digital tools, yet there is little guidance in these culturally focused studies on how AI might be incorporated in ways that preserve, not dilute, cultural specificity.

This disconnect presents a critical challenge. Many indigenous African communities are multilingual, low-resource, and digitally underserved. Introducing AI in these contexts raises serious questions about data ownership, ethical consent, and long-term accessibility. Technologies developed in global centers may not account for local norms, oral traditions, or even the symbolic meanings embedded in language use. Without frameworks for ethical and participatory technology deployment, well-intentioned AI solutions risk becoming extractive or ineffective. Thus, the present study positions itself at the intersection of cultural integrity and technological innovation. Rather than seeing AI as a universal fix, it investigates how it can be harnessed responsively, adapting to the linguistic diversity, infrastructural realities, and social priorities of African communities engaged in language preservation.

In sum, while traditional approaches to documentation and revitalization remain foundational, they are increasingly insufficient in the face of modern linguistic threats and generational shifts. At the same time, AI technologies, while promising, must be critically adapted, not merely applied, within African contexts to ensure they support rather than disrupt local efforts. This study advances the conversation by asking not only what AI *can* do for language preservation, but under what conditions and with what community input it *should* be used.

2.2. Digital Archiving of Endangered Languages

Efforts to preserve endangered languages have increasingly intersected with digital innovation, particularly through the use of Artificial Intelligence (AI). Recent studies, including those by Wong et al. [21], emphasize the transformative role of AI in automating traditionally labor-intensive aspects of language documentation. Through the use of machine learning and natural language processing (NLP), these tools can expedite transcription, data collection, and linguistic analysis, functions that historically required extensive time and expertise. Similar claims are echoed in the work of Jarrahi et al. [22], where automation is framed as a solution to the resource constraints plaguing traditional linguistic fieldwork.

While such studies successfully showcase the technical efficiencies of AI, they often treat automation as a standalone achievement, missing deeper questions around context-specific applicability and ethical deployment. For instance, although AI can process massive volumes of audio data with minimal human intervention, this advantage is not evenly distributed. In multilingual African communities with low digital literacy or inadequate infrastructure, the gains of automation may remain largely theoretical. More importantly, these approaches risk divorcing documentation from the social and cultural practices that sustain language vitality. Treating documentation as a purely technical task overlooks the need for speaker agency, informed consent, and community participation, elements crucial to ensuring that technology supports, rather than supplants, human-centered language preservation.

Contrasting this technically focused lens, researchers delve into the use of AI-powered speech recognition for transcribing unwritten, low-resource languages [23,24]. In contexts where literacy is low and linguistic expertise scarce, these tools offer a compelling solution. They allow oral languages to be captured, transcribed, and analyzed in ways that facilitate not only preservation but also revitalization and education. However, their effectiveness varies significantly across African contexts. Many AI models, originally trained on well-resourced languages, falter when exposed to tonal distinctions, morphologically rich structures, or dialectal variation, features common in African languages. These failures reveal a broader issue: the assumption that AI tools are universally applicable often ignores the linguistic diversity and sociocultural nuance inherent to indigenous language communities.

Even when such systems work technically, their deployment raises critical socio-ethical challenges. Issues of data privacy, algorithmic bias, and unequal access to the outputs of these technologies are often underexplored. As

Amiri highlights, speech-to-text tools are often celebrated for their speed and accuracy [23], but little attention is paid to who owns the resulting data, how it is stored, or whether communities have the right to withdraw or control access to their linguistic heritage. These omissions can reinforce historical patterns of extractive research, especially when data collected from marginalized groups is archived or published without meaningful reciprocity or consent.

Studies extend the conversation to corpus generation and linguistic modeling, showing how machine learning can be applied to segment speech, detect patterns, and build annotated datasets [25,26]. While these tools expand the scope of linguistic analysis, they also depend on large, clean datasets, something rarely available in the African language landscape. Moreover, many endangered languages are oral and context-bound, meaning that meaning shifts with speaker, setting, or tone. Such nuances are not easily captured by computational models, which tend to generalize in ways that may flatten or distort cultural meaning. The reliance on models trained in high-resource environments leads to algorithmic brittleness when applied to African contexts, revealing a persistent gap between technical innovation and cultural applicability.

This disconnect is especially stark in the domain of digital archiving. Studies recognize the promise of AI-enhanced archiving systems, which can automate metadata tagging, facilitate content indexing, and streamline data retrieval [27,28]. These platforms theoretically allow endangered languages to be stored, shared, and accessed more efficiently. However, while these systems are technologically robust, their practical relevance in many African communities is questionable. Endangered language speakers often live in regions with poor connectivity, limited access to digital tools, and minimal training in how to use such platforms. In such settings, even the best-designed archiving systems can fail to reach their intended audience.

Moreover, these studies frequently underplay the ethical concerns of digitizing and archiving linguistic materials. Once a language is digitized and uploaded to a cloud server or academic repository, questions arise: Who controls this content? Can communities withdraw or alter it? What happens when sacred or restricted speech forms become globally accessible? These questions are particularly urgent in African contexts, where oral traditions often include knowledge that is socially or spiritually bounded. Current AI archiving frameworks rarely build in mechanisms for community control, potentially alienating the very people they aim to support.

Critically, while many of the reviewed studies showcase what AI *can* do for documentation and archiving, few interrogate what it *should* do, or how. There is a recurring tendency to treat technology as a neutral tool rather than a socially embedded practice. As a result, important questions about linguistic justice, digital sovereignty, and community empowerment remain unanswered. Simply inserting AI into existing documentation workflows is not enough; these technologies must be adapted to fit the cultural, infrastructural, and ethical realities of African language communities.

The current study builds on this critique by examining how AI-driven archiving platforms can be deployed in ways that are not only efficient but also inclusive, accountable, and culturally grounded. Rather than seeing AI as a universal solution, the focus here is on co-creating technologies with communities, respecting indigenous knowledge systems, and ensuring that digital preservation enhances, not replaces, human relationships with language.

2.3. Challenges in Implementing AI Driven Documentation of Endangered Languages

The rapid evolution of Artificial Intelligence (AI) technologies has opened new pathways for language preservation, particularly in regions where linguistic diversity is both a cultural asset and a vulnerability. Among the most notable advancements is Neural Machine Translation (NMT), which, as Martin argues [29], has outpaced traditional statistical and rule-based methods in generating more fluent and context-aware translations. For African languages, many of which are underrepresented in digital spaces, this offers the potential to increase visibility, support intercommunity communication, and embed indigenous languages into educational and online ecosystems.

Yet, these promises are tempered by deep structural limitations. NMT systems rely heavily on large, high-quality datasets, and for most indigenous African languages, such resources remain scarce. As Martin concedes [29], the lack of expansive, annotated corpora results in degraded

translation quality, particularly for languages with rich morphology or non-standardized orthographies. Moreover, the study fails to address the ethical terrain accompanying these technical gaps. Without clear frameworks around data ownership or strategies for inclusive representation, NMT risks privileging certain dialects or dominant varieties at the expense of linguistic minorities, potentially replicating patterns of digital exclusion under the guise of innovation.

Wasike et al. further expand on these technical challenges, pointing out that tonal variation, inflectional complexity, and dialectal fragmentation, features common across African linguistic contexts, are often flattened in AI models optimized for more rigid grammatical structures [30]. These linguistic intricacies are not easily captured through standard training pipelines, and their omission can result in systems that output grammatically plausible but culturally or semantically inaccurate translations. What's missing from much of the current literature is a critical interrogation of how AI systems can be made responsive to this diversity, beyond simply scaling existing models. The current study builds on this by exploring alternative strategies for training AI on limited data, including community-generated corpora and low-shot learning techniques tailored to typologically diverse languages.

At the same time, Nekoto et al. offer an optimistic view of progress [31], highlighting growing efforts to integrate African languages into the broader AI ecosystem. The development of language datasets, bespoke translation tools, and localization frameworks represents a shift from mere technological importation to more inclusive language tech development. However, even here, the discussion tends to overstate progress without accounting for the scalability and sustainability of these efforts. Many of these initiatives remain pilot projects, funded externally and disconnected from local educational or policy infrastructures. More importantly, the question of who drives these projects—tech firms, academic institutions, or communities themselves—is rarely interrogated. Without centering indigenous agency, even well-meaning innovations risk reproducing extractive dynamics that prioritize model performance over cultural relevance or linguistic justice.

What Adebara makes particularly clear is that these technological limitations are inseparable from broader digital inequities [32]. Rural and remote communities, where linguistic diversity is often highest, are frequently those with the weakest access to digital infrastructure. This digital divide extends beyond hardware or connectivity; it includes disparities in digital literacy, limited access to training for using AI tools, and a general absence of localized interfaces or documentation in indigenous languages. The result is that even when AI tools are theoretically available, their practical deployment remains concentrated in urban centers or research institutions. Moreover, without community-informed protocols for data governance, consent, and access, the digitization of language risks becoming a form of cultural dispossession rather than preservation.

While these challenges are acknowledged across the literature, a deeper problem persists: existing research rarely provides empirical evidence on how AI tools actually perform in context, on the ground, in real communities, with real speakers. Eglash reinforces this point by noting that most studies fail to evaluate user experiences, trust in AI systems, or the social dynamics that affect tool uptake [33]. This oversight is especially problematic in African contexts, where language is deeply intertwined with identity, spirituality, and oral tradition. Ignoring these layers in favor of purely technical assessments leads to tools that may function well in testing environments but falter in meaningful use.

What is needed, and what the current study addresses, is a shift in perspective: from evaluating AI as a set of universal technologies to understanding it as a culturally situated practice. This involves not only adapting tools to fit linguistic diversity but also aligning them with local knowledge systems, ethical norms, and community priorities. Success in AI-driven language preservation cannot be measured by translation accuracy or dataset size alone; it must also be judged by who benefits, who decides, and who controls the means of preservation.

3. Theoretical Framework

The study is anchored in the Appropriating Technology Theory, originally propounded by Carroll and Fidock [34]. This theory examines how marginalized and non-dominant communities actively adapt, reinterpret, and repurpose technology to meet their own unique cultural, social, and political needs, rather than passively adopting technolo-

gies developed externally. This framework is especially pertinent to the current study, which investigates the role of Artificial Intelligence (AI) in the preservation and revitalization of indigenous African languages.

Applying this theory, the study posits that AI-driven language revitalization efforts should be tailored to align closely with the cultural and linguistic specificities of local communities. Rather than viewing technology as a neutral, one-size-fits-all solution, Appropriating Technology Theory emphasizes that successful technological interventions are those shaped by the values, knowledge systems, and social structures of the communities they intend to serve.

For example, AI tools such as speech recognition software, automated transcription systems, and neural machine translation models should be co-developed with the active involvement of native speakers, local linguists, cultural custodians, and community members. This participatory approach ensures that the technology is culturally sensitive, linguistically accurate, and responsive to the diverse dialects and usage patterns inherent in African languages. Moreover, it fosters community ownership and empowerment, which are critical for sustainable language preservation.

This theory has been applied in various contexts to demonstrate the power of local agency in technology use. In the context of AI and language preservation, Appropriating Technology Theory aligns with contemporary concerns in AI ethics and sustainability [35]. It highlights the importance of respecting data sovereignty, ensuring informed consent, and protecting cultural intellectual property when developing AI tools for indigenous languages. Furthermore, it calls for sustainable development practices that prioritize long-term community benefits over short-term technological gains. By grounding the study in Appropriating Technology Theory, the research advocates for a model of AI integration that is not only technically effective but also ethically responsible and culturally sustainable. This approach recognizes the importance of community-driven innovation in preserving linguistic heritage, ensuring that AI technologies empower rather than marginalize indigenous language speakers.

4. Materials and Method

This study employs a systematic review methodol-

ogy to examine the role of artificial intelligence (AI) in the documentation and revitalization of indigenous African languages. The systematic review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines according to [36], ensuring a structured and transparent approach to data collection, analysis, and synthesis. This research design was chosen to synthesize existing research on AI-driven language documentation and revitalization. This approach allows for a comprehensive evaluation of peer-reviewed studies, identifying key trends, challenges, and opportunities in the field.

4.1. Data Sources and Selection Criteria

The study sourced data from twenty academic digital repositories and peer-reviewed journals related to computational linguistics, AI applications in language preservation, and African linguistics. The following four databases were searched: Google Scholar, SpringerLink, ScienceDirect, and African Journals Online (AJOL)

4.1.1. Search Strings

The following Boolean search strings were used across all databases to identify relevant peer-reviewed publications:

- "Artificial intelligence" AND "language documentation" AND Africa
- "Neural machine translation" AND "indigenous African languages"
- "Speech recognition" AND "low-resource languages" AND Africa
- "Digital archiving" AND "endangered languages" AND "AI"
- "AI-driven transcription" AND "African linguistic diversity"
- "Machine learning" AND "language revitalization" AND "Africa"

4.1.2. Date Range

The search focused on studies published between January 2020 and March 2025, a period selected to capture the most recent advances in AI and their application to African language preservation.

4.1.3. Inclusion Criteria

Studies published between 2020 and March 2025 to capture recent advancements in AI technology. Research focusing on AI-driven documentation, transcription, machine translation, and digital archiving of indigenous languages. Studies applying AI methodologies such as speech recognition, neural machine translation (NMT), and corpus development.

4.1.4. Exclusion Criteria

Studies that do not specifically address language preservation. Research that focuses on non-AI-based language preservation methods. Non-peer-reviewed articles, opinion pieces, and grey literature.

4.1.5. Screening Process and Study Selection

The initial search yielded 612 studies across all four databases:

- Google Scholar: 295
- ScienceDirect: 121
- SpringerLink: 98
- AJOL: 98

After removing 113 duplicates, 499 titles and abstracts were screened for relevance. Following the application of inclusion and exclusion criteria, 87 full-text articles were assessed in detail. Out of these, 35 studies met all eligibility criteria and were included in the final synthesis.

4.1.6. Data Collection and Analysis

A PRISMA flowchart was used to track the selection process, ensuring transparency in study inclusion and exclusion. Extracted data included: study objectives, AI methods applied, language(s) studied, key findings, and challenges noted. Thematic analysis was conducted to identify recurring themes such as AI-driven transcription, language modeling, digital archiving, and ethical considerations.

4.2. Ethical Considerations

The study adhered to ethical research standards, ensuring that all reviewed studies were properly cited and respected intellectual property rights. AI-driven language revitalization was considered, highlighting the need for ethical AI development and data ownership.

5. Results and Discussions

Figure 1 shows the frequency of key themes identified in the 20 reviewed literature sources.

Each bar represents the relative frequency of a theme in the systematic review which form part of discussion, with AI-powered speech recognition taking the lead.
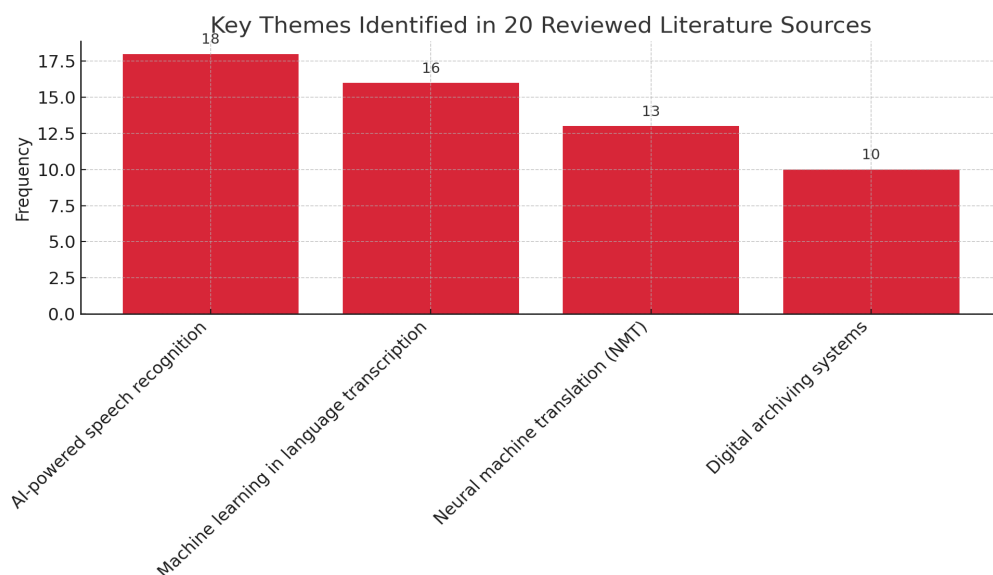

Figure 1. AI Mechanisms Potential for Revitalizing and Documenting African Languages.

5.1. AI-Powered Language Documentation and Revitalization

The reviewed literature affirms that AI technologies, particularly those based on advanced speech recognition models, hold substantial potential to enhance the automatic transcription of languages, especially in low-resource language contexts. Through supervised machine learning, these models are trained to detect patterns in spoken language by aligning audio data with transcribed text. This enables systems to segment audio streams, identify phonemes, map them to words, and generate increasingly accurate written transcriptions. When applied effectively, AI can curate large-scale repositories of linguistic content in African languages, encompassing lexicons, oral stories, proverbs, and grammatical rules, through the combined power of speech recognition and natural language processing (NLP). In this way, AI acts not merely as a transcription tool but as a curatorial medium, systematically organizing indigenous language resources for long-term preservation, analysis, and revitalization.

An illustrative example is drawn from *Dholuo*, a Nilotic language spoken by the Luo community in Kenya:

(1) Dholuo proverb: *Ka ineno remb wuon gi, to pok inenori*

AI translation: *If you see your father's blood, you haven't yet seen your own.*

This example demonstrates AI's capacity not only to transcribe but to semantically tag proverbs according to themes such as wisdom, inheritance, and moral instruction. In such cases, AI systems can also enrich metadata by identifying usage contexts and speaker demographics, enhancing the cultural depth and linguistic richness of digital archives. Furthermore, AI could support the creation of searchable, morphologically aware dictionaries, linking entries like *paro* (to think) to relevant verb conjugation patterns and examples drawn from oral narratives. These capabilities align with findings by Wong et al. [21], who observe that AI has accelerated the automation of core language documentation processes, particularly through supervised learning that internalizes phonological structures and intonational patterns unique to each language.

Supporting this, Martin explains that machine learning approaches can be fine-tuned to segment continuous speech into meaningful linguistic units and generate annotated corpora [29], resources that are essential for downstream tasks such as grammar modeling, educational tool development, and digital dictionary construction. These corpora serve not only as repositories of linguistic data but also as building blocks for community-centered language revitalization initiatives. However, the analysis also reveals critical limitations in the current application of AI to indigenous African languages, most notably, data scarcity. As emphasized by Adebara [32], many African languages are under-documented and lack the volume of annotated, high-quality datasets needed to train robust AI models. This scarcity limits both the accuracy and the generalizability of AI systems. For instance, without sufficient data covering dialectal variation or tonal differentiation, models often produce transcriptions that are inaccurate or culturally disconnected, especially in languages that are predominantly oral or lack standardized orthographies. The challenge is further compounded by the complexity of African linguistic systems, which often include non-linear morphological structures, extensive tone marking, and regionally specific expressions that are difficult to capture using conventional NLP frameworks.

This issue is consistent with the concerns raised by Ajani et al. [18], who highlight the vital role of documentation in preserving linguistic diversity and reinforcing cultural identity. Yet, AI models that overlook these linguistic and sociocultural nuances risk generating outputs that are technically functional but contextually flawed. Moreover, reliance on general-purpose or Western-centric training data can perpetuate linguistic inequities by marginalizing less dominant dialects or regional speech forms. While the integration of AI, particularly supervised learning and NLP, has demonstrably advanced the capacity for automated transcription and linguistic data curation, its full potential in African contexts remains constrained by infrastructural and data limitations. To support equitable language preservation, future AI development must focus on collaborative corpus-building, the inclusion of diverse dialectal input, and culturally informed model training strategies. Without such measures, AI systems risk replicating the very marginalization they aim to resolve.

Appropriating Technology Theory provides a grounded and culturally responsive lens through which AI solu-

tions for language documentation can be more thoughtfully designed. It encourages moving beyond technical feasibility to consider how people use, adapt, and assign meaning to technology in their own terms. By applying this theory, AI developers and researchers can build tools that do more than function. They can be welcomed, trusted, and sustained by the communities they aim to serve. As a result, community-based data collection initiatives and collaborations with native speakers become essential for developing linguistically and culturally accurate datasets. While AI-powered transcription offers transformative potential for the preservation and revitalization of indigenous languages, its effectiveness depends on deliberate efforts to expand language resources, incorporate local linguistic expertise,

and design inclusive AI frameworks that are culturally sensitive and contextually grounded.

5.2. Digital Archiving and Corpus Development

The reviewed literature underscores a growing emphasis on integrating Artificial Intelligence (AI) mechanisms in the revitalization and documentation of endangered and under-resourced languages (**Table 1**). This trend aligns with the broader goals of digital archiving and corpus development, particularly in constructing reliable linguistic resources for future use. From the literature reviewed, the presented data indicate level of focus on AI mechanisms in language revitalization and documentation.

Table 1. Areas of Focus Revitalization and Documentation of African Languages.

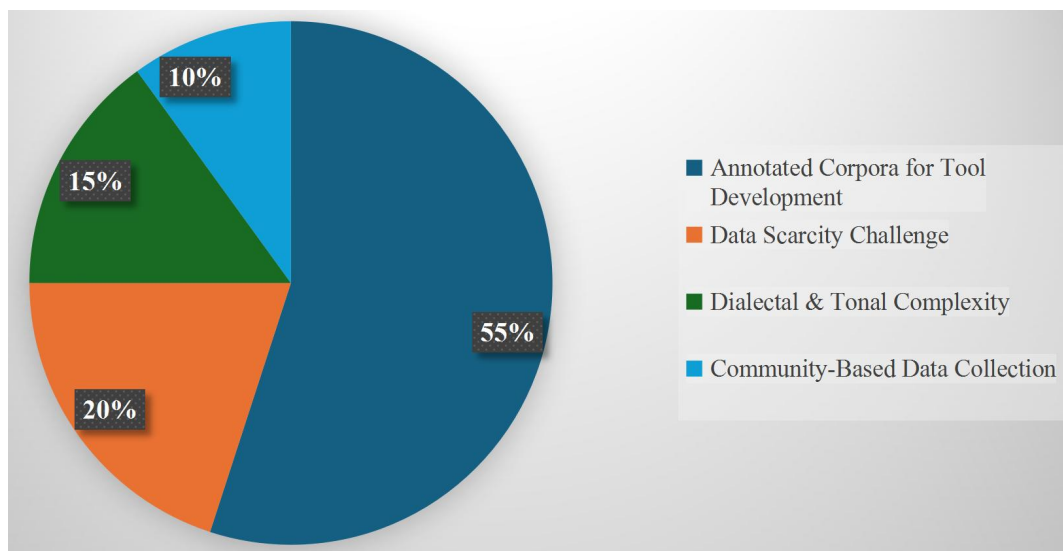| Area of Focus / Concern	Percentage (%)
Annotated Corpora for Tool Development	55%
Data Scarcity Challenge	20%
Dialectal & Tonal Complexity	15%
Community-Based Data Collection	10%

This data is presented in **Figure 2**.

Figure 2. Empirical Studies Focusing on AI-Powered Language Documentation and Revitalization.

Empirical data from the reviewed literature indicate that annotated corpora development is the most prominent focus in AI-driven language documentation, accounting for 55% of the studies. This dominant trend highlights

the foundational role that annotated corpora play in training AI systems such as speech recognition models, neural machine translation (NMT) engines, and grammar parsers, all central components of digital archiving infrastructures.

As [26] confirms, annotated corpora are crucial for enabling AI applications to function effectively in endangered language contexts, providing the structured linguistic input necessary for accurate parsing, transcription, and semantic analysis. Visual representation of the reviewed studies supports this observation by showing a concentration of research and development activity in corpus construction, illustrating a clear prioritization of technical infrastructure over community-centered methods.

In contrast, data scarcity, representing 20% of the studies, emerges as a persistent barrier to progress. Many indigenous African languages lack the large-scale, high-quality datasets required to train robust AI models, a limitation that impairs the performance and adaptability of language technologies across diverse linguistic contexts. This aligns with the findings of Wasike et al. [30], who discuss how the absence of extensive digital corpora constrains the capacity of NMT systems to learn accurate, context-sensitive translations. While this challenge is frequently acknowledged, it is often addressed in general terms, with few studies providing granular strategies for overcoming it, an analytical gap that should be addressed in future work. Moreover, the visual mapping of this theme may underestimate the complexity of the issue, as some studies mention data scarcity only briefly, meaning it might not be fully captured during keyword-based coding.

Another critical issue is the dialectal and tonal complexity of African languages, emphasized in 15% of the studies. Many endangered languages exhibit phonetic variation, tonal depth, and regional dialects, which present unique challenges to AI systems trained primarily on standardized, morphologically simpler languages. According to O'Shaughnessy [25], advanced machine learning algorithms can be tailored to detect and learn from such variability, but doing so requires data collection methods that preserve rather than flatten linguistic diversity. The representation of this theme in the literature suggests growing awareness of these challenges, yet practical implementation remains uneven, and training data often lacks fine-grained metadata about dialectal and tonal features, another limitation of current corpus-building efforts.

Community-based data collection, although comprising only 10% of the studies, is perhaps the most ethically and culturally significant component. This low representa-

tion highlights a systemic imbalance in AI-focused language documentation: while technical models are heavily developed, the active participation of indigenous communities, particularly native speakers, remains minimal. Involving local communities in data collection, transcription, and annotation not only enriches the quality of the data but also ensures cultural authenticity, ethical accountability, and long-term sustainability. Failure to do so risks perpetuating extractive research practices and undermines the very goals of language preservation. This concern aligns with broader critiques regarding cultural appropriation, especially when indigenous data is collected and utilized without benefit-sharing or informed consent.

AI technologies are undeniably capable of enhancing data collection, annotation, and digital preservation. By leveraging machine learning, NLP, and speech recognition, AI can capture and process vast amounts of linguistic data at speeds and scales unattainable through manual methods. A compelling example is seen in the transcription of a Dholuo sentence, spoken by the Luo community in Kenya:

(2) Dholuo: *Wuon dala odhi e piny*

AI transcription: *The head of the home has gone to the village*

This example demonstrates how AI can be used not only to transcribe spoken language but also to tag metadata such as speaker age, dialectal variation, and sociolinguistic context. When aggregated across large datasets, such tagging enables the creation of searchable, metadata-rich corpora that preserve both the linguistic structure and the cultural nuance of endangered languages. These applications align with the findings of Amiri [23], who emphasizes the effectiveness of AI-driven transcription tools in supporting educational materials, language learning platforms, and revitalization efforts.

Platforms such as the Endangered Languages Archive (ELAR) and Mozilla Common Voice exemplify how AI technologies are being mobilized in large-scale preservation efforts. ELAR serves as a digital repository for uploading and accessing audiovisual language documentation, while Mozilla Common Voice employs a crowdsourcing model to collect voice samples for training open-source speech recognition engines. These initiatives have significantly contributed to the growth of annotated corpora, lexicographic development, and digital language

learning tools, as confirmed by [27], who highlight the scalability and efficiency of such platforms. However, a critical limitation persists: the limited involvement of the communities whose languages are being documented. In many cases, native speakers and cultural custodians are excluded from the design, governance, and benefit-sharing aspects of these tools, raising serious ethical concerns. The visual analysis and literature synthesis reveal a dominant focus on corpus-building and technical model development, with far less emphasis placed on community collaboration and ethical engagement. While AI offers powerful tools for documenting and revitalizing endangered African languages, its success ultimately depends on addressing issues of data quality, linguistic complexity, and inclusive design. Without a stronger commitment to community-led and context-aware development, AI risks reinforcing existing disparities rather than supporting truly equitable and sustainable language preservation.

This raises important questions about data sovereignty and ownership, particularly when linguistic resources are stored, processed, or disseminated by external institutions without transparent agreements or informed consent. Appropriating Technology Theory (ATT) helps to illuminate these challenges by emphasizing that communities do not passively adopt technologies; they actively reshape and recontextualize them based on their own cultural values, priorities, and constraints. When AI systems are developed without meaningful community input, they risk being rejected, underutilized, or viewed with distrust, regardless of their technical sophistication.

To ensure that AI-driven language preservation efforts are ethical, respectful, and sustainable, it is essential to move beyond extractive or top-down approaches and instead adopt community-centered frameworks rooted in appropriation rather than imposition. This involves not only involving indigenous speakers in data collection and annotation processes but also granting them agency and control over how their linguistic data is stored, accessed, and applied. It also means designing AI tools that are adaptable to local contexts, technologies that can be shaped by community needs rather than forcing communities to conform to external technical standards.

ATT further suggests that collaborative partnerships grounded in local knowledge systems are essential for fostering trust and ensuring cultural legitimacy. When communities are empowered to co-direct the development of digital archives and language technologies, they are more likely to appropriate these tools in meaningful and enduring ways. Moreover, ensuring that the tangible benefits of preservation, such as educational resources, revitalization programs, and increased language visibility, are returned to the communities themselves reinforces ethical reciprocity and long-term engagement. By applying Appropriating Technology Theory, developers and researchers are reminded that success in digital language preservation is not only measured by technical outputs but by how well these technologies align with, and are adapted to, the social, ethical, and cultural landscapes of the communities they intend to serve.

5.3. Challenges in AI-driven Language Revitalization

Despite the promising potential of AI in supporting language documentation and revitalization as highlighted by Wang [10], who argues that recent advancements in AI technologies offer innovative pathways for preserving endangered languages, several significant challenges hinder its widespread and equitable implementation, particularly in the context of indigenous African languages. While AI tools such as speech recognition, neural machine translation (NMT), and corpus annotation have introduced new levels of efficiency, the realities on the ground reveal deep structural barriers that must be critically addressed for these technologies to succeed meaningfully and ethically.

One of the most pressing obstacles is data scarcity, a challenge consistently cited across the literature and confirmed by Wasike et al. [30], who emphasize that the effectiveness of machine learning models is heavily constrained by limited training data and linguistic complexity. Most AI systems, especially those based on supervised learning, depend on large volumes of annotated text and audio data to accurately model grammar, phonology, and semantics. However, many indigenous African languages remain under-documented, lacking the foundational linguistic resources, such as transcribed texts, lexicons, and high-quality audio corpora, necessary to train such systems effectively. In this study, we observe that this limitation not only reduces model performance but also severely im-

pairs the scalability and adaptability of AI systems across dialects and linguistic domains. The underrepresentation of these languages in AI training data often leads to models that are linguistically inaccurate, culturally insensitive, or simply inapplicable in real-world community contexts.

Beyond technical limitations, ethical concerns present another layer of complexity. The use of AI in language documentation involves the collection, processing, and dissemination of cultural and linguistic data, much of which is deeply embedded in local traditions, belief systems, and oral knowledge. Without robust ethical frameworks, there is a real risk of extractive practices, where data is gathered from indigenous communities without informed consent, proper attribution, or reciprocal benefits. This concern is particularly urgent in African contexts where historical patterns of marginalization and exploitation persist. Data ownership, community agency, and benefit-sharing mechanisms must be central to the design of AI systems. Furthermore, cultural representation within AI models must be approached with care. A model that misclassifies a ceremonial expression or fails to capture the social significance of a proverb can inadvertently distort or erase vital elements of linguistic identity, thus undermining the goals of revitalization rather than supporting them.

A further and often underemphasized challenge is technological accessibility. As noted by Adebara [32], many African communities continue to face a significant digital divide, characterized by poor internet infrastructure, limited access to computing devices, and low levels of digital literacy. These barriers mean that even when AI tools are developed, their practical deployment within target communities is not guaranteed. Moreover, most AI platforms and tools are designed with users from high-resource settings in mind, often failing to consider local hardware constraints, language interfaces, or user experience design suited to rural or low-literacy environments. This disconnect between tool design and user context often renders even the most advanced AI models inaccessible or irrelevant to those they are intended to support.

Together, these challenges underscore the urgent need for context-sensitive, ethically grounded, and participatory approaches to AI in language documentation and revitalization. Technological innovation alone is insufficient; what is required is a shift in both research and implementation paradigms, one that recognizes indigenous communities not merely as data sources or passive beneficiaries but as active co-creators and custodians of their linguistic heritage. Addressing issues of data scarcity, ethical practice, and accessibility holistically will be critical to developing AI systems that are not only effective but also inclusive, equitable, and sustainable in their impact.

These factors significantly hinder the ability of local speakers and researchers to engage meaningfully with AI tools and platforms, thereby limiting grassroots participation in language preservation efforts.

Appropriating Technology Theory (ATT) helps explain this challenge by emphasizing that technologies are not simply adopted; they are reshaped and reinterpreted through local cultural, social, and linguistic contexts. When AI systems are introduced without accounting for these realities, they often fail to resonate with or serve the needs of the communities they target. A key technical barrier is the bias embedded in AI models, which are typically trained on data from high-resource, structurally simpler languages and thus struggle to generalize to the low-resource, morphologically complex, and tonal languages prevalent across Africa. This leads to inaccurate transcriptions, misclassifications, and the continued marginalization of African languages within digital ecosystems. From an ATT perspective, this reflects a misalignment between the design of AI tools and the contexts in which they are intended to operate. To address this, it is essential to develop tailored models and context-specific training approaches that are co-created with local communities, respect linguistic diversity, and support flexible adaptation. By embedding local knowledge and user agency into the development process, AI technologies can be more effectively appropriated, ensuring accurate, equitable, and culturally grounded language preservation solutions.

6. Conclusions and Recommendations

Artificial Intelligence (AI) presents transformative opportunities for the documentation and revitalization of indigenous African languages. Technologies such as automated transcription, neural machine translation, and digital archiving have the potential to significantly enhance the efficiency and scale of language preservation efforts.

These tools can help overcome traditional barriers related to resource constraints and linguistic expertise, enabling broader and more systematic recording and revitalization of endangered languages.

However, the successful application of AI in this domain is contingent upon addressing several critical challenges. These include the scarcity of annotated linguistic data necessary to train effective AI models, ethical concerns related to community agency, cultural representation, and data sovereignty, as well as technological inequities resulting from limited digital access, particularly in rural and marginalized communities. Additionally, biases embedded in AI systems trained predominantly on high-resource languages risk marginalizing indigenous languages further if not conscientiously mitigated.

To ensure that AI-driven language preservation is both effective and inclusive, interdisciplinary and community-driven approaches are essential. Integrating local knowledge, cultural norms, and linguistic nuances into the design and deployment of AI tools fosters relevance and sustainability. This underscores the importance of collaboration between linguists, AI developers, indigenous language speakers, and cultural custodians.

6.1. Recommendations

Collaborative Development: AI tools for language preservation should be co-created with indigenous communities, ensuring cultural sensitivity and contextual appropriateness.

Capacity Building: Investments in digital literacy and technological training within indigenous populations are vital to promote local ownership and sustainability of AI applications.

Ethical Practices: Establish comprehensive ethical frameworks that respect data sovereignty, ensure informed consent, and protect intellectual property rights of indigenous communities.

Addressing Data Scarcity: Prioritize initiatives to build annotated corpora and develop AI models that perform well in low-resource language contexts, accommodating linguistic diversity.

Bridging the Digital Divide: Implement infrastructural and policy measures to improve digital access in underserved regions to facilitate equitable participation in AI-driven language preservation.

Ongoing Research: Support empirical studies that evaluate AI's impact on language vitality, community perceptions, and the sociocultural implications of technology adoption.

National Integrated Language Policy: This policy should be based on linguistic human rights, cultural sustainability, and the integration of modern technology to support documentation, revitalization, and intergenerational transmission.

Funding

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Informed consent was obtained from all subjects involved in the study.

Data Availability Statement

The data supporting the findings of this study are available from the corresponding author upon reasonable request. They are not publicly available due to privacy and ethical considerations.

Acknowledgments

Conflicts of Interest

The author declares no conflict of interest.

References

[1] Ndoleriire, O.K., 2024. The Development and intellectualisation of African languages revisited. Africa Journal of Language Studies. 1(1), 1–33.

[2] Kemei, J.N., Lusambili, K.M., Okoth, P.G., et al., 2023. History of Community and indigenous language mass media evolution in Turkana County from

1963–2022. Athens Journal of History. 9, 1–19. DOI: https://doi.org/10.30958/ajhis.X-Y-Z

[3] Ihejirika, C., 2024. Harnessing African indigenous knowledge systems for knowledge production: A re-definition of a culture-centric epistemology. Journal of Contemporary Philosophical and Anthropological Studies. 2(1), 3–14. DOI: https://doi.org/10.59652/jcpas.v2i1.103

[4] Awal, A., 2024. Endangered languages: A systematic qualitative study of socio-cultural impacts and revitalization. Darnioji Daugiakalbystė. 25, 65–101.

[5] Adamou, E., 2024. Endangered languages. MIT Press: Cambridge, MA, USA.

[6] Ogwudile, C.E.C., 2024. An exploratory study of mother tongu and language endengerment process. Journal of Modern European Languages And Literatures. 18(2), 26–36.

[7] Zulkiflee, Z., Chuchu, S.I.F.H., 2025. Belait language maintenance in Brunei Darussalam. International Journal of Linguistics and Literature. 3(1), 77–92. DOI: https://doi.org/10.36312/ijlic.v3i1.2617

[8] Zhong, T., Yang, Z., Liu, Z., et al., 2024. Opportunities and challenges of large language models for low-resource languages in Humanities Research. arXiv. arXiv:2412.04497. 9 December 2024. DOI: https://doi.org/10.48550/arXiv.2412.04497

[9] Jermakowicz, E., 2023. The coming transformative impact of large language models and artificial intelligence on lobal business and education. Journal of Global Awareness. 4(2), 1–22. DOI: https://doi.org/10.24073/jga/4/02/03

[10] Wang, L., 2024. Artificial intelligence's role in the realm of endangered languages: Documentation and teaching. Applied and Computational Engineering. 48, 123–129. DOI: https://doi.org/10.54254/2755-2721/48/20241249

[11] Rehm, G., Way, A., 2023. Strategic research, innovation and implementation agenda for digital language equality in Europe by 2030. In: Rehm, G., Way, A. (eds.). European language equality: A strategic agenda for digital language equality. Springer International Publishing: Cham, Switzerland. pp. 387–412. DOI: https://doi.org/10.1007/978-3-031-28819-7_45

[12] Llanes-Ortiz, G., 2023. Digital initiatives for indigenous languages. UNESCO Publishing: Paris, France.

[13] Dennis, N.K., 2024. Using AI powered speech recognition technology to improve English pronunciation and speaking skills. IAFOR Journal of Education. 12(2), 107–126.

[14] Xuan, Q., Yang, Y., 2024. New developments in English teaching and translation methods in thec-Converged media environment: An AI-based analysis. Intelligent Systems & Robotic Mechanics. 1(1), 1–10.

[15] Meek, B.A., 2022. "At risk" languages and the road to recovery: A case from the Yukon. Journal of Multilingual and Multicultural Development. 43(3), 228–242. DOI: https://doi.org/10.1080/01434632.2022.2050381

[16] Machin-Mastromatteo, J.D., 2023. Community-driven and social initiatives. Information Development. 39(3), 393–401. DOI: https://doi.org/10.1177/02666669231197243

[17] Tohit, N.F.M., Haque, M., 2024. Preparing the younger generation for an aging society: Strategies, challenges, and opportunities. Cureus. 16(7), e64121. DOI: https://doi.org/10.7759/cureus.64121

[18] Ajani, Y.A., Oladokun, B.D., Olarongbe, S.A., et al., 2024. Revitalizing indigenous knowledge systems via digital media technologies for sustainability of indigenous languages. Preservation, Digital Technology & Culture. 53(1), 35–44. DOI: https://doi.org/10.1515/pdtc-2023-0051

[19] Mgimwa, P.A., Dash, S.R., 2024. Reviving endangered languages: Exploring AI technologies for the preservation of Tanzania's Hehe language. In: Mohanty, S.S., Dash, S.R., Parida, S. (eds.). Applying AI-based tools and technologies towards revitalization of indigenous and endangered languages. Springer Nature: Singapore. pp. 23–33. DOI: https://doi.org/10.1007/978-981-97-1987-7_2

[20] Alaimo, C., Kallinikos, J., 2022. Organizations decentered: Data objects, technology and knowledge. Organization Science. 33(1), 19–37. DOI: https://doi.org/10.1287/orsc.2021.1552

[21] Wong, M.-F., Guo, S., Hang, C.-N., et al., 2023. Natural language generation and understanding of big code for AI-assisted programming: A Review. Entropy. 25(6), 888. DOI: https://doi.org/10.3390/e25060888

[22] Jarrahi, M.H., Askay, D., Eshraghi, A., et al., 2023. Artificial intelligence and knowledge management: A partnership between human and AI. Business Horizons. 66(1), 87–99. DOI: https://doi.org/10.1016/j.bushor.2022.03.002

[23] Amiri, S.M.H., 2025. Beyond language barriers: Multilingual NLP and voice recognition for global connectivity. International Journal of Science and Research Archive. 15(2), 406–419. DOI: https://doi.org/10.2139/ssrn.5254434

[24] Jarvenpaa, S.L., Essén, A., 2023. Data sustainability: Data governance in data infrastructures across technological and human generations. Information and Organization. 33(1), 100449. DOI: https://doi.org/10.1016/j.infoandorg.2023.100449

[25] O'Shaughnessy, D., 2024. Trends and developments in automatic speech recognition research. Computer Speech & Language. 83, 101538. DOI: https://doi.org/10.1016/j.csl.2023.101538

[26] Incelli, E., 2025. Exploring the future of corpus lin-

guistics: Innovations in AI and social impact. International Journal of Mass Communication. 3, 1–10. DOI: https://doi.org/10.6000/2818-3401.2025.03.01

[27] Ingram, M., 2025. The Role of AI in language preservation and revitalization. In: Ardalan, I.D., Banifatemi, A., Gonzalez, F., et al. (eds.). AI for Community. Chapman and Hall/CRC: Boca Raton, FL, USA. pp. 55–80.

[28] Oyighan, D., Ukubeyinje, E.S., David-West, B.T., et al., 2024. The role of AI in transforming metadata management: Insights on challenges, opportunities, and emerging trends. Asian Journal of Information Science and Technology. 14(2), 20–26. DOI: https://doi.org/10.70112/ajist-2024.14.2.4277

[29] Martin, S., 2024. Advancements in neural machine translation: Techniques and applications. Journal of Innovative Technologies. 7(1), 1–9.

[30] Wasike, A., Kamukama, I., Abass, Y.A., et al., 2024. Advancements in natural language understanding-driven machine translation: Focus on English and the low resource dialectal Lusoga. International Journal of Innovative Science and Research Technology. 9(10), 470–480. DOI: https://doi.org/10.38124/ijisrt/IJISRT24OCT410

[31] Nekoto, W., Marivate, V., Matsila, T., et al., 2020. Participatory research for low-resourced machine translation: A case study in African languages. arXiv. arXiv:2010.02353. 6 November 2020. DOI: https://doi.org/10.48550/arXiv.2010.02353

[32] Adebara, I., 2024. Towards Afrocentric natural language processing [PhD Thesis]. University of British Columbia: Vancouver, Canada. DOI: https://doi.org/10.14288/1.0440415

[33] Eglash, R., 2004. Appropriating Technology: Vernacular Science and Social Power. University of Minnesota Press: Minneapolis, MN, USA.

[34] Carroll, J., Fidock, J., 2011. Beyond resistance to technology appropriation. In Proceedings of the 44th Hawaii International Conference on System Sciences, Kauai, HI, USA, 04-07 January 2011; pp. 1–9. DOI: https://doi.org/10.1109/HICSS.2011.82

[35] Page, M.J., McKenzie, J.E., Bossuyt, P.M., et al., 2021. Updating guidance for reporting systematic reviews: Development of the PRISMA 2020 statement. Journal of Clinical Epidemiology. 134, 103–112. DOI: https://doi.org/10.1016/j.jclinepi.2021.02.003

[36] Arya, S., Kaji, A. H., & Boermeester, M. A. (2021). PRISMA reporting guidelines for meta-analyses and systematic reviews. JAMA surgery, 156(8), 789-790.