Zamfara International Journal of Education (ZIJE)
The Official Journal of the Faculty of Education
Federal University Gusau, Zamfara State, Nigeria
Volume 6, Number 6, October, 2026
pISSN: (Hard Copy): 2814 – 1377
eISSN (Online): 2814 – 1369

www.foefugusau.com.ng
Reviewed Date: 14-10-2026
Publication Date: 31-10-2026
DOI: **https://doi.org/10.64348/zije.2026648**

Leveraging Technology and Translation in Archiving and Propagating The history of Yoruba Language in Nigeria

OMOLE, Cecilia Chinedu
omole.cecilia1979@fceokene.edu.ng
08033741936
Department of French, Federal College of Education, Okene, Kogi State, Nigeria

Abstract

The Yoruba language is one of the major African languages and an important carrier of the history, identity, traditions, philosophy, oral literature, and cultural values of the Yoruba people. However, much of Yoruba cultural and linguistic knowledge has historically been transmitted orally, while a significant proportion of older written, audio, and audiovisual materials remain scattered across libraries, private collections, religious institutions, museums, and family archives. The rapid development of digital technology provides opportunities to preserve, organize, access, and disseminate these materials for present and future generations. This study examines the role of technology and translation in archiving and propagating the history of the Yoruba language. It focuses on digital repositories, digitization, electronic dictionaries, speech technology, machine translation, optical character recognition, online learning platforms, and social media as tools for preserving and promoting Yoruba. Recent research demonstrates growing resources for Yoruba natural language processing, including parallel translation corpora, speech datasets, and dialectal resources. However, significant challenges remain, particularly inadequate digitization of historical materials, limited high-quality datasets, orthographic and diacritic issues, dialectal variation, copyright concerns, and weaknesses in culturally sensitive machine translation. The study argues that technological preservation should not merely store Yoruba materials but should make them searchable, translatable, educationally useful, and accessible to Yoruba-speaking communities. It recommends collaboration among linguists, historians, translators, technologists, libraries, government institutions, cultural organizations, and native speakers to establish sustainable digital Yoruba archives.

Keywords: Yoruba language, language preservation, translation, technology & machine translation.

Introduction

Language is an important repository of the history and cultural identity of a people. Through language, communities preserve knowledge about their origins, traditions, beliefs, social institutions, historical events, occupations, arts, literature, and worldview. Bamgbose (1969) notes that Yoruba has a long history of scholarly study and a substantial body of literature, demonstrating the importance of the language as a vehicle for documenting and preserving Yoruba knowledge and culture. For the Yoruba people of West Africa, the Yoruba language (Èdè Yorùbá) performs this cultural and historical function through oral traditions, poetry, proverbs, songs, praise poetry, myths, folktales, historical narratives, religious texts, written literature, and everyday communication. Yoruba is also a tonal language with several regional varieties, making its documentation important not only for communication but also for the preservation of regional identities and cultural knowledge. Ahia et al. (2024) observe that Yoruba encompasses a continuum of regional dialects and that differences in vocabulary, pronunciation, spelling, and syntax reflect regional, social, and cultural differences. The history of Yoruba has been preserved through both oral and written forms. Oral forms such as àlọ́, òwe, ìjálá, oríkì, songs, chants, and traditional narratives contain information about Yoruba society and its worldview. Bamgbose (1969) explains that the development of Yoruba studies has been supported by a substantial body of literature produced over more than a century. This written tradition has complemented the older oral systems through which Yoruba historical and cultural knowledge was transmitted. The documentation of Yoruba has also developed through linguistic scholarship. Bamgbose (1969), for example, contributed to the study and standardization of Yoruba, including work associated with Yoruba orthography. His work illustrates the importance of establishing written conventions through which a language can be taught, recorded, published, and preserved. Nevertheless, historical and cultural materials are not always available in easily accessible and sustainable formats. Some materials exist only as printed books, handwritten documents, recordings, photographs, manuscripts, or privately held collections. Such materials can deteriorate, become difficult to access, or disappear when they are not systematically documented and preserved. The need for digital preservation is therefore particularly important for languages whose cultural heritage is distributed across written and oral sources. Digital technology has changed the

Zamfara International Journal of Education (ZIJE)
The Official Journal of the Faculty of Education
Federal University Gusau, Zamfara State, Nigeria
Volume 6, Number 6, October, 2026
pISSN: (Hard Copy): 2814 – 1377
eISSN (Online): 2814 – 1369
www.foefugusau.com.ng
Reviewed Date: 14-10-2026
Publication Date: 31-10-2026
DOI: https://doi.org/10.64348/zije.2026648

possibilities for language documentation and cultural preservation. Digitization allows books, manuscripts, photographs, audio recordings, videos, dictionaries, newspapers, and oral histories to be converted into digital formats. These materials can then be stored in digital repositories, searched electronically, copied for preservation, and made accessible to researchers, students, and communities. UNESCO (2024) emphasizes the importance of digital initiatives in preserving, revitalizing, and promoting Indigenous and underrepresented languages. Digital technologies can also support the development of the datasets required for automatic speech recognition, machine translation, and other language technologies. The application of technology to Yoruba is becoming increasingly significant. Ògúnrẹ̀mí et al. (2024) developed the ÌròyìnSpeech corpus, a Yoruba speech dataset designed to support Text-to-Speech (TTS) and Automatic Speech Recognition (ASR). The researchers curated approximately 23,000 Yoruba text sentences and used participatory approaches, including crowdsourcing through Mozilla Common Voice, to increase the availability of Yoruba speech data. This development demonstrates how technology can assist not only in creating modern language applications but also in preserving spoken Yoruba for future generations.

Technology is also being used to address the problem of dialectal underrepresentation. Ahia et al. (2024) developed YORULECT, a parallel text and speech corpus covering four regional Yoruba varieties. Their study found substantial performance differences between Standard Yoruba and regional dialects in machine translation, automatic speech recognition, and speech-to-text translation. However, they also demonstrated that dialect-adaptive training could reduce some of these differences. This finding suggests that Yoruba archiving should not concentrate exclusively on Standard Yoruba but should also document regional varieties and their associated cultural knowledge. Translation is another important means of propagating Yoruba history and knowledge. Translation can make Yoruba historical, literary, and cultural materials accessible to people who do not speak Yoruba. At the same time, translating knowledge from English and other languages into Yoruba can expand the functional use of Yoruba in education, science, technology, administration, literature, and digital communication. Translation involving Yoruba presents linguistic and cultural challenges. Adebara, Abdul-Mageed, and Silfverberg (2022) demonstrate that differences between Yoruba and English can create ambiguity in machine translation. In particular, features that are expressed morphologically in English may be expressed contextually in Yoruba, creating difficulties for automated translation systems. Their study compared statistical and neural machine-translation approaches and showed that different systems performed differently depending on the linguistic category being translated. The challenge is even greater when translation involves culturally specific materials such as proverbs, oríkì, traditional songs, folktales, religious expressions, and historical narratives. Such materials often contain meanings that cannot be adequately captured through direct word-for-word translation. Consequently, effective Yoruba translation requires attention to linguistic structure, cultural context, tone, idiomatic expressions, and the intended meaning of the original text. Recent research further indicates that Yoruba remains comparatively under-resourced in Natural Language Processing (NLP). Jimoh, De Wille, and Nikolov (2025), in a systematic review of 105 Yoruba NLP studies published between 2014 and 2024, identify the scarcity of annotated corpora, limited availability of pretrained language models, tonal complexity, and dependence on diacritics as important challenges. They nevertheless report growing development of both monolingual and multilingual Yoruba resources.

The development of these technologies therefore creates an important opportunity for Yoruba language preservation. Digital archives can preserve historical documents; speech corpora can preserve spoken language; machine translation can facilitate access to Yoruba materials across languages; and artificial intelligence can support transcription, search, classification, language learning, and other forms of digital dissemination. Therefore, the archiving and propagation of Yoruba history should involve an integrated approach in which digital preservation, language documentation, translation, education, community participation, and modern language technologies work together. The combination of these approaches can help ensure that Yoruba historical knowledge is not only preserved but also made accessible and useful to present and future generations.

Statement of the Problem

Zamfara International Journal of Education (ZIJE)
The Official Journal of the Faculty of Education
Federal University Gusau, Zamfara State, Nigeria
Volume 6, Number 6, October, 2026
pISSN: (Hard Copy): 2814 – 1377
eISSN (Online): 2814 – 1369
www.foefugusau.com.ng
Reviewed Date: 14-10-2026
Publication Date: 31-10-2026
DOI: https://doi.org/10.64348/zije.2026648

The preservation and transmission of Yoruba language and cultural heritage face several challenges. Many historical materials remain undigitized and difficult to access, while oral traditions are threatened by globalization, language shift, and reduced intergenerational transmission. Yoruba is also underrepresented in language technology, with limited digital resources and inadequate representation of regional dialects. In addition, machine translation may produce inaccurate or culturally inappropriate translations because of Yoruba's tonal system, idioms, cultural expressions, and grammatical differences from English. Therefore, there is a need for effective integration of digital archiving, technology, translation, and community participation to preserve and propagate the history and cultural heritage of the Yoruba language for present and future generations.

Objectives of the Study

The general objective of this study is to examine how technology and translation can be used to archive, preserve, and propagate the history and cultural heritage of the Yoruba language.

The study seeks to:

1. Examine the importance of archiving Yoruba linguistic and historical materials.
2. Identify technological tools that can be used to preserve Yoruba language and cultural heritage.
3. Examine the role of digitization in preserving Yoruba books, manuscripts, photographs, audio recordings, videos, and oral traditions.

Literature Review

Concept of Language Archiving

Language archiving refers to the systematic collection, preservation, organization, storage, and provision of access to materials that document a language. According to Berez-Kroeker, Gabber, and Slayton (2023), modern language documentation has become closely connected with digital technologies that enable researchers and communities to create, store, share, and mobilize language resources. They emphasize that language documentation increasingly involves not only linguistic description but also the creation of recorded language events, annotations, and digital resources that can be used for language learning, revitalization, research, and community purposes. Archiving therefore goes beyond preserving grammar books and dictionaries. It may include written texts, dictionaries, photographs, audio recordings, videos, interviews, conversations, songs, stories, oral histories, and linguistic analyses. Zaušková (2022), in her discussion of linguistic legacy materials, argues that older language documentation materials are valuable resources for contemporary digital archives but often present challenges relating to formats, metadata, ethical considerations, and proper curation. She emphasizes that effective digital archiving requires systematic organization and appropriate infrastructure so that both historical and contemporary materials can remain usable. Similarly, Babinski et al. (2022) argue that the existence of a digital language collection does not automatically guarantee that the materials will be useful to researchers or communities. Their review found problems relating to the discoverability, accessibility, and usability of digital language collections. This suggests that language archives should not merely store materials but should also provide adequate transcripts, metadata, organization, and access mechanisms that enable users to find and understand the archived resources.

For Yoruba, language archiving should therefore encompass both written and oral forms of linguistic and cultural expression. Historical books, newspapers, manuscripts, dictionaries, photographs, traditional stories, oríkì, òwe (proverbs), songs, religious texts, philosophical materials, interviews, and recordings of native speakers can all constitute important Yoruba archival resources. Recent Yoruba research demonstrates the practical importance of such digital resources. Ògúnrẹ̀mí et al. (2024), for example, developed the Ìròyìn Speech corpus to increase the availability of high-quality Yoruba speech data for Automatic Speech Recognition (ASR) and Text-to-Speech (TTS). Their project curated approximately 23,000 text sentences and produced about 42 hours of speech recorded by 80 volunteers, together with additional validated recordings collected through Mozilla Common Voice. This demonstrates how contemporary technology can transform spoken Yoruba into a structured digital resource that can support both language technology and long-term preservation. According to Ogunyale (2024) provides a particularly relevant example of cultural archiving through A Documentation of Onko/Yoruba Traditional Communication Media. The collection contains recordings and interviews documenting traditional communication practices in communities in Oyo and Osun States. Such documentation illustrates how digital

Zamfara International Journal of Education (ZIJE)
The Official Journal of the Faculty of Education
Federal University Gusau, Zamfara State, Nigeria
Volume 6, Number 6, October, 2026
pISSN: (Hard Copy): 2814 – 1377
eISSN (Online): 2814 – 1369
www.foefugusau.com.ng
Reviewed Date: 14-10-2026
Publication Date: 31-10-2026
DOI: https://doi.org/10.64348/zije.2026648

archives can preserve not only the linguistic form of Yoruba but also the cultural practices and communication systems associated with the language.

Technology and Digital Preservation

Digital technology provides important tools for preserving cultural materials in formats that can be copied, searched, indexed, and distributed. Digitization can involve scanning printed materials, photographing manuscripts, converting analogue audio to digital formats, transcribing speech, and creating metadata that describes each resource. UNESCO (2019) notes that the use of information and communication technologies to promote linguistic diversity contributes to "the preservation, access and promotion of the invaluable traditional knowledge imbued in languages." For Indigenous languages, UNESCO and Global Voices emphasize the potential of digital technologies to "conserve, revive, and promote Indigenous languages" (Llanes-Ortiz & Huiza Capo, 2023). The Digital Initiatives for Indigenous Languages toolkit, published in 2023, further identifies approaches including facilitating, multiplying, normalizing, educating, reclaiming, imagining, defending, and protecting Indigenous languages in digital spaces. For Yoruba, digitization can therefore make historical materials more accessible to researchers, students, and younger generations. A properly designed digital repository could allow users to search materials by author, date, location, dialect, subject, genre, speaker, or historical period.

Yoruba Language and Natural Language Processing

Natural Language Processing (NLP) involves computational techniques for processing human language. It supports applications such as spell-checking, search engines, text classification, speech recognition, machine translation, question answering, and conversational systems. Yoruba NLP has expanded considerably in recent years. Researchers have developed datasets and models for translation, speech recognition, text-to-speech, and other language tasks. A significant development is the MENYO-20k corpus, which provides a multidomain Yoruba-English parallel corpus with carefully curated orthography and standardized evaluation splits. Such resources are important because machine translation requires large quantities of high-quality parallel text. Similarly, Ìròyìn Speech was developed as a contemporary Yoruba speech corpus for text-to-speech and automatic speech recognition. The project included approximately 23,000 text sentences and recordings from volunteers, demonstrating the value of participatory approaches to Yoruba speech-data creation.

Technology and Yoruba Dialects

One important issue in Yoruba language technology is dialectal diversity. Computational systems trained mainly on standardized Yoruba may perform less effectively when processing regional varieties. According to Ahia, Aremu, Abagyan, Gonen, Adelani, Abolade, Smith, and Tsvetkov (2024) developed YORULECT, a corpus covering four regional Yoruba dialects and three domains. Their findings showed substantial performance differences between standard Yoruba and regional dialects, while dialect-adaptive fine-tuning helped reduce these disparities. This finding is important for historical archiving because dialects contain valuable information about regional history and cultural identity. Digital preservation should therefore not focus exclusively on standardized Yoruba.

Translation and the Yoruba Language

Translation facilitates communication between Yoruba speakers and speakers of other languages. It can also support the international dissemination of Yoruba literature, history, philosophy, oral traditions, and cultural knowledge. Machine translation has become an important area of research because neural and multilingual models can translate between languages with limited computational resources. However, Yoruba-English translation presents specific linguistic challenges. Adebara, Abdul-Mageed, and Silfverberg (2022) examined Yoruba-English machine translation and found that differences in the way grammatical information is expressed can create ambiguity during translation. Similarly, Adegbehingbe, Soronnadi, Adebara, and Adekanmbi (2025) found that multilingual models face challenges in specialized Yoruba translation, particularly because of vocabulary mismatch, data scarcity, and morphological complexity. Consequently, machine translation should complement rather than completely replace human translators, especially for historical, literary, religious, legal, and culturally sensitive Yoruba materials.

Speech Technology

Zamfara International Journal of Education (ZIJE)
The Official Journal of the Faculty of Education
Federal University Gusau, Zamfara State, Nigeria
Volume 6, Number 6, October, 2026
pISSN: (Hard Copy): 2814 – 1377
eISSN (Online): 2814 – 1369
www.foefugusau.com.ng
Reviewed Date: 14-10-2026
Publication Date: 31-10-2026
DOI: https://doi.org/10.64348/zije.2026648

Speech technology is particularly valuable for archiving oral history. Automatic Speech Recognition (ASR) can convert Yoruba speech into written text, while Text-to-Speech (TTS) can convert written Yoruba into spoken language. These technologies can make Yoruba materials more accessible to people with different literacy levels and can support language learning. However, Yoruba's tonal system, dialectal diversity, pronunciation differences, and culturally specific vocabulary create challenges for automatic speech systems. Recent work presented at Africa NLP highlights problems in recognizing culturally embedded Yoruba oral literature, including proverbs, and emphasizes the need for more culturally representative speech datasets.

Research Gap

The major research gap is that existing Yoruba language technology research is largely fragmented**,** focusing on computational tasks such as machine translation, speech recognition, and corpus development rather than the integrated preservation of Yoruba historical and cultural heritage**.** Current datasets also have limited coverage of Yoruba dialects, historical periods, social groups, and cultural domains, while historical materials are underrepresented. Furthermore, archiving and translation are rarely integrated, and computational evaluation often fails to adequately measure the preservation of cultural meaning. There is also limited attention to sustainable, community-centered digital repositories that address ownership, privacy, cultural sensitivity, metadata, accessibility, and long-term preservation. Therefore, the key research gap is the absence of a comprehensive, culturally informed digital framework that integrates Yoruba historical archiving, digitization, transcription, translation, metadata, dialect representation, community validation, and long-term cultural preservation. This study seeks to address this gap by treating archiving, language technology, translation, historical preservation, and cultural propagation as interconnected elements of Yoruba language preservation**.**

Conclusion

Yoruba language preservation requires a combination of digital technology and human cultural knowledge. Although advances in speech technology, machine translation, NLP, and digital resources provide valuable opportunities, effective preservation also depends on accurate data, native-speaker participation, dialect representation, cultural protection, translation quality, and sustainable digital archives. Therefore, digitizing historical materials, recording oral traditions, developing culturally appropriate language technologies, and creating accessible repositories can help preserve and propagate Yoruba heritage for future generations in Nigeria and across the global Yoruba diaspora.

Recommendations

Based on the discussion, the study recommends the following:

1. Establishment of Yoruba Digital Archives: Universities, libraries, museums, cultural organizations, and government agencies should collaborate to establish centralized and interoperable digital repositories for Yoruba historical and linguistic materials.
2. Digitization of Historical Materials: Old Yoruba books, manuscripts, newspapers, photographs, recordings, and other historical documents should be systematically digitized and preserved.
3. Development of Yoruba Corpora: More high-quality Yoruba text, speech, and parallel translation corpora should be created to support NLP, machine translation, speech recognition, and artificial intelligence.
4. Preservation of Dialects: Digital projects should document and represent regional Yoruba dialects rather than concentrating exclusively on standardized Yoruba.
5. Improvement of Machine Translation: Researchers should develop Yoruba-English and English-Yoruba translation systems using culturally relevant datasets and human evaluation.
6. Human Validation of AI Translation: Machine-generated translations of proverbs, historical documents, literature, religious texts, and culturally sensitive materials should be reviewed by qualified Yoruba speakers, linguists, historians, and translators.

References

Adebara, I., Abdul-Mageed, M., & Silfverberg, M. (2022). Linguistically-motivated Yorùbá-English machine translation. *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*. Association for Computational Linguistics. (ACL Anthology)

Zamfara International Journal of Education (ZIJE)
The Official Journal of the Faculty of Education
Federal University Gusau, Zamfara State, Nigeria
Volume 6, Number 6, October, 2026
pISSN: (Hard Copy): 2814 – 1377
eISSN (Online): 2814 – 1369

www.foefugusau.com.ng
Reviewed Date: 14-10-2026
Publication Date: 31-10-2026
DOI: **https://doi.org/10.64348/zije.2026648**

Adelani, D. I., Ruiter, D., Alabi, J. O., Adebonojo, D., Ayeni, A., Adeyemi, M., Awokoya, A. E., & España-Bonet, C. (2021). The effect of domain and diacritics in Yoruba-English neural machine translation. *Proceedings of Machine Translation Summit XVIII: Research Track*, 61–75.

Ahia, O., Aremu, A., Abagyan, D., Gonen, H., Adelani, D. I., Abolade, D., Smith, N. A., & Tsvetkov, Y. (2024). Voices unheard: NLP resources and models for Yorùbá regional dialects. *Proceedings of EMNLP 2024*. Association for Computational Linguistics.

Ahia, O., Kreutzer, J., & Hooker, S. (2021). The low-resource double bind: An empirical study of pruning for low-resource machine translation. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3316–3333.

Llanes-Ortiz, G., & Huiza Capo, A. (2023). *Digital initiatives for Indigenous languages*. UNESCO.

Nwafor, E., & Andy, A. (2022). A survey of machine translation tasks on Nigerian languages. *Proceedings of LREC 2022*. Association for Computational Linguistics.

Ógúnrẹ̀mí, T., Tùbọ̀sún, K., Aremu, A., Orife, I., & Adelani, D. I. (2024). ÌròyìnSpeech: A multi-purpose Yorùbá speech corpus. *Proceedings of LREC-COLING 2024*, 9296–9303.

Signoroni, E., & Rychlý, P. (2023). Evaluating sentence alignment methods in a low-resource setting: An English-Yorùbá study case. *Proceedings of the Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages*, 123–129.

UNESCO. (2024). *Digital preservation of Indigenous languages: At the intersection of technology and culture*. UNESCO. (UNESCO)

Yorlect. (2026). *Advancing Yoruba language technology: Harnessing Yoruba dialects in technology*. (yorlect.org)