International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Development of a Multilingual Translation System for Nigeria Language

Samuel Ifeanyi Ojukwu*[1], Ikechukwu Ekene Onyenwe[2], Ebele G. Onyedinma[3]

[1,2,3] Department of Computer Science, Nnamdi Azikiwe University, Awka, Anambra State, Nigeria
samuel_ojukwu@yahoo.com[1], ie.onyenwe@unizik.edu.ng[2], eg.osita@unizik.edu.ng[3]

ABSTRACT

Nigeria is one of the most linguistically diverse countries in the world, with over 500 indigenous languages spoken across its regions. Despite this rich linguistic heritage, many Nigerian languages remain underrepresented in digital tools and resources, limiting equitable access to information and opportunities. This research presents the design and development of a multilingual translation system specifically tailored for Nigerian languages. The study began with the collection of English textual data and engaged linguistic experts to create high-quality language resources for selected indigenous languages. To address challenges of data scarcity and translation accuracy, the system employs Natural Language Toolkit (NLTK) methods to tokenize text into sentences and words, enabling expert translators to provide precise equivalents in their native languages. Additionally, the system integrates a similarity checker and human evaluation module to ensure the quality and consistency of translations. Initial performance tests indicate that the similarity checker achieves an effectiveness rate of 80% to 98.5% for straightforward translations. By supporting resource creation and enabling multilingual information access, this system contributes to the preservation and promotion of Nigeria's linguistic diversity and lays a foundation for future Natural Language Processing (NLP) applications for under-resourced languages.

Keywords: Sentiment Analysis, Information Retrieval, Named Entity Recognition, Text Categorization, Text Summarization.

1.0 Introduction

Nigeria is one of the most linguistically diverse countries in the world, with over 525 indigenous languages spoken by more than 250 ethnic groups (Afolabi, 2023). This extraordinary linguistic richness represents a vital part of Nigeria's cultural heritage and identity. However, many of these languages face the threat of endangerment and extinction due to factors such as rapid urbanization, globalization, migration, and the dominance of a few widely used languages (Agantiem, 2017). A language that is not preserved or supported digitally loses its speakers over time, resulting in the erosion of unique cultural knowledge and identity embedded within that language. In today's increasingly interconnected world, multilingual natural language processing (NLP) plays a critical role in bridging language barriers and providing equitable access to information and services (Krishna, 2023). NLP enables machines to process, understand, and generate human languages and supports various applications such as machine translation, speech recognition, sentiment analysis, text classification, information retrieval, and summarization. These applications are now indispensable for digital communication, education, governance, and social inclusion. However, most African languages, including most Nigerian indigenous languages, remain severely under-resourced and are excluded from the benefits of modern NLP technologies (Eberhard, Gary, Simons, & Fennig, 2023). Nigeria's major languages: Hausa, Igbo, and Yoruba are widely spoken in the North, South-East, and South-West respectively (Olajoke & Idowu, 2013). Yet, even these dominant languages lack the extensive computational resources available for high-resource languages like English or Chinese. The situation is even more challenging for the hundreds of other indigenous languages that have little to no digital representation. This lack of NLP resources severely limits speakers' access to online information, educational materials, and technology in their native languages, deepening the digital divide and threatening cultural preservation. Addressing this gap requires the development of scalable frameworks that can support the creation of robust NLP tools for under-resourced Nigerian languages. This includes building high-quality bilingual datasets, designing effective translation pipelines, and involving linguistic experts to ensure cultural and linguistic accuracy. While global companies have made significant progress in machine translation for major world languages, little effort has been dedicated to the unique needs of Nigeria's diverse linguistic communities. This research introduces the Development of a Multilingual Translation System for Nigerian Languages. The system is designed to tackle the challenges of data scarcity, translation accuracy, and linguistic diversity by combining modern NLP techniques with local linguistic expertise. The approach begins with the collection of English textual content and uses the Natural Language Toolkit (NLTK) to tokenize this content into sentences and words. Linguistic experts then provide accurate translations into selected Nigerian languages. A similarity checker and human evaluation processes are integrated to maintain high translation quality and consistency. By establishing a practical and adaptable workflow for building new bilingual textual resources, this system lays the groundwork for future NLP applications, including neural machine translation, speech-to-text systems, sentiment analysis, and conversational AI tailored to Nigeria's linguistic realities.

Ultimately, this research aims to help bridge the digital divide, promote digital inclusiveness, and contribute to the preservation and revitalization of Nigeria's rich linguistic and cultural heritage in the digital era.

2.0 Related Works

Multilingual natural language processing (NLP) has attracted growing attention as researchers and practitioners work to bridge global language barriers and promote equitable access to digital content (Krishna, 2023). Multilingual NLP systems power essential applications including machine translation, language modeling, text classification, sentiment analysis, information retrieval, and summarization. However, for many African and specifically Nigerian indigenous languages, the lack of digitized corpora and robust NLP tools remains a major challenge (Eberhard, Gary, Simons, & Fennig, 2023). In the Nigerian context, notable progress has been made. Onyenwe (2017) developed pioneering NLP resources to advance Igbo NLP research. This work produced the first Igbo NLP toolkit and introduced a new part-of-speech (POS) tagset, IgbTS, adapted from the EAGLES guidelines to better reflect Igbo's unique linguistic features. Onyenwe's work also proposed a monolingual-based manual annotation bootstrapping approach using inter-annotation agreement outcomes and an affix-based tag-error correction method to improve the tagging of morphologically inflected words. These contributions marked a critical step toward establishing a Basic Language Resource Kit (BLARK) for the Igbo language. Fagbolu et al. (2015) created an electronic corpus database that supports natural language processing and machine translation. Their project utilized mobile and web programming tools, including JDK, Android SDK, PHP, MySQL, .NET, and related technologies, to build accessible platforms for language data collection and validation. Abdullahi (2016) designed a multilingual translation system to support agricultural e-extension services in Nigeria, applying a hybrid method combining rule-based and statistical machine translation to deliver research outputs in English to farmers in Arabic, Hausa, Igbo, and Yoruba. The system achieved an accuracy rate of 65% in translating technical information to local languages. Ayegba, Musa, and Philip (2016) developed an Igala-to-English machine translation system using a rule-based approach, tested on 250 Igala texts. The system achieved an accuracy of 81.2%, evaluated through human assessment. Demas (2016) built an Amharic-English translation app for iOS devices, integrating Microsoft's Translator API and a customized Amharic keyboard. Although functional, the system did not match the quality of human translation, revealing ongoing challenges for low-resource languages. Beyond Nigeria, Hou and Li-Hong (2022) proposed an interactive machine translation (IMT) framework using segment analysis and online knowledge sources to improve translation quality through human-computer interaction. Despite its innovation, corpus testing showed persistent discourse-level errors. Desai and Dabhi (2022) provided a detailed survey of resource and component development for Gujarati NLP systems, evaluating prominent tools and identifying gaps that hinder scalability for low-resource languages. Choudhary (2021) documented India's national initiative through the Linguistic Data Consortium for Indian Languages (LDC-IL), which focuses on systematic linguistic data creation for 20 indigenous languages to advance NLP research and applications. In domain-specific NLP, Muller, Salathé, and Kummervold (2023) developed COVID-Twitter-BERT (CT-BERT), a transformer-based model trained on COVID-19-related tweets. CT-BERT demonstrated the impact of domain-adapted NLP models for tasks like classification and question answering in specialized contexts. Khurana, Koli, Khatter, and Singh (2022) presented a comprehensive discussion on the evolution of NLP, key developments in Natural Language Generation, emerging trends, common datasets, and current evaluation metrics.

Conclusively, these studies highlight global and local initiatives to expand language resources and translation systems for under-resourced and endangered languages. Despite this progress, Nigeria still lacks scalable multilingual translation frameworks that combine linguist validation, robust text tokenization, and quality assurance checks in a single system. This research builds on existing efforts by developing an adaptable multilingual translation system tailored for Nigeria's indigenous languages. By combining expert translation input, NLTK-based text processing, and a built-in similarity checker with human review, the system provides a practical approach for creating and validating new NLP resources that help preserve Nigeria's linguistic diversity in the digital age.

3.0 Materials and Methods

This study followed a structured design and implementation methodology, integrating natural language processing (NLP) techniques, human linguistic expertise, and user-friendly software tools to develop a practical multilingual translation system for Nigerian indigenous languages. The development process included system design, data preparation, translation workflow, quality assurance, and system evaluation.

3.1 System Flow and Components

The system architecture was designed as shown in **Figure 1** (System Flowchart) and includes the following core components and stages:

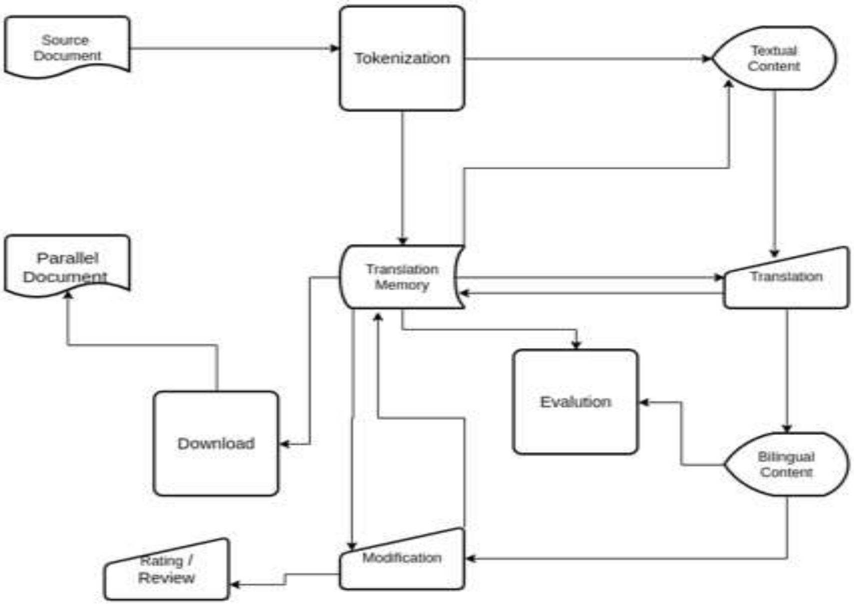


Figure 1: System Flowchart

- **Source Document:** The process begins with the upload of raw English textual files. These documents are drawn from diverse sources such as news articles, stories, grammar books, dictionaries, and textbooks.
- **Tokenization:** Using the Natural Language Toolkit (NLTK), the system automatically tokenizes the uploaded raw text into sentences and words. This segmentation enables precise alignment for translation and facilitates the creation of a parallel corpus.
- **Textual Content Display:** The tokenized sentences and words are displayed in the system's translator interface, where bilingual experts translate them into the target Nigerian languages.
- **Translation:** Linguists input accurate translations for each sentence or word, leveraging their cultural and contextual knowledge to ensure high-quality output.
- **Translation Memory:** Each translated pair (source and target) is stored in a translation memory database using **SQLite**, enabling efficient reuse of verified translations and supporting future NLP resource development.
- **Parallel Document Generation:** The system can generate and download parallel documents containing both the original English text and its translated version.
- **Download Module:** A built-in download function allows users to store the bilingual output securely on local devices for backup or further processing.
- **Bilingual Content Display:** The system visually presents the source English text alongside its translation for reference, review, or export.
- **Evaluation Module:** A similarity checker automatically assesses the alignment between the source and translated text, providing a similarity score to flag possible inconsistencies.
- **Rating/Review:** Human reviewers conduct additional quality control by rating and commenting on each translation, ensuring cultural appropriateness and linguistic accuracy.
- **Modification:** Translators can revise their submissions based on reviewer feedback, ensuring continuous improvement of the bilingual dataset.

3.2 Dataset and Preparation

A diverse English textual dataset was carefully curated to ensure linguistic richness and practical relevance. Texts were selected to cover a range of vocabulary, sentence structures, and contexts. Bilingual speakers and linguistic experts collaborated to translate the segmented text into Nigeria's three major languages: Hausa, Igbo, and Yoruba. The segmentation step divided large volumes of text into individual sentences and words to create manageable units for building the parallel dataset. Multiple bilingual speakers reviewed the translated text to verify meaning preservation and contextual accuracy. To complement this manual process, the system's similarity checker provided automated feedback by calculating similarity scores for each translation pair.

3.3 Experimental Tools and Technologies

The system was implemented using a modern web development stack optimized for NLP tasks:

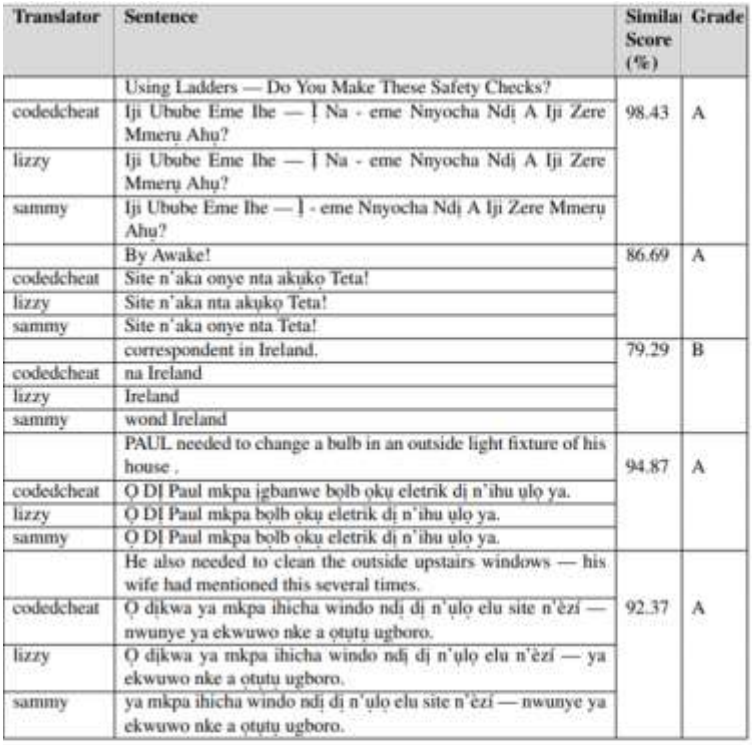
- **Python (v12):** An open-source programming language with rich libraries and frameworks. Core Python libraries used included **NLTK** for text tokenization, **Django** for building the web-based user interface, and **Scikit-learn** for evaluating basic NLP operations.
- **SQLite:** A lightweight relational database management system used to store the tokenized segments, translation pairs, reviewer comments, and system configurations.
- **Tailwind CSS:** A modern front-end framework providing pre-designed HTML, CSS, and JavaScript components to create an intuitive, responsive, and visually appealing interface for translators and reviewers.
- **Similarity Checker:** Custom Python scripts were integrated to calculate the semantic similarity between the source and translated texts to guide reviewers and maintain translation quality.

All modules were tested on a Windows 10 development environment, ensuring that the system remains accessible and lightweight for use by linguists and researchers with limited technical resources.

4.0 Results and Discussion

To evaluate the multilingual translation system, sample English sentences were translated into Igbo by three translators: Codedcheat, Lizzy, and Sammy. The translations were assessed using the system's automated similarity checker and verified by human reviewers who assigned grades based on meaning, fluency, and accuracy. Table 1 illustrates how well the system supported consistent translations across different users. Similarity scores ranged from **79.29% to 98.43%**, showing strong alignment between the source text and the translations. Higher scores, such as **98.43%** for "Using Ladders - Do You Make These Safety Checks?" and **94.87%** for "PAUL needed to change a bulb…" received **A grades**, confirming the effectiveness of the tokenization, display, and translation memory modules. Even the lowest score (**79.29%**) was graded **B**, indicating that the essential meaning was preserved despite minor differences in word choice or phrasing. The consistency of scores and grades across multiple translators demonstrates that the system's workflow; from tokenizing sentences to storing them in the translation memory works as intended. The similarity checker effectively flags quality levels, while human grading helps refine outputs, especially for phrases involving proper nouns or idiomatic expressions. The presence of the modification and review modules allows translators to adjust and improve translations before final storage. Overall, these results confirm that the system reliably produces high-quality bilingual text, bridging the resource gap for Nigerian languages. This test validates the system's practical value for building reusable parallel datasets to support machine translation and other NLP tasks for under-resourced languages. Thus, the similarity scores confirm that the translation system is reliable for creating reusable bilingual datasets for under-resourced Nigerian languages. The positive agreement between automated and human assessments underscores its value for building foundational resources for future machine translation and NLP applications.

TABLE I: Sample Data Evaluation using Similarity Checker Test



Translator	Sentence	Simila Score (%)	Grade
	Using Ladders — Do You Make These Safety Checks?	98.43	A
codedcheat	Iji Ubube Eme Ihe — Ị Na - eme Nnyocha Ndị A Iji Zere Mmerụ Ahụ?		
lizzy	Iji Ubube Eme Ihe — Ị Na - eme Nnyocha Ndị A Iji Zere Mmerụ Ahụ?		
sammy	Iji Ubube Eme Ihe — Ị - eme Nnyocha Ndị A Iji Zere Mmerụ Ahụ?		
	By Awake!	86.69	A
codedcheat	Site n'aka onye nta akụkọ Teta!		
lizzy	Site n'aka nta akụkọ Teta!		
sammy	Site n'aka onye nta Teta!		
	correspondent in Ireland.	79.29	B
codedcheat	na Ireland		
lizzy	Ireland		
sammy	wond Ireland		
	PAUL needed to change a bulb in an outside light fixture of his house .	94.87	A
codedcheat	Ọ DỊ Paul mkpa ịgbanwe bọlb ọkụ eletrik dị n'ihu ụlọ ya.		
lizzy	Ọ DỊ Paul mkpa bọlb ọkụ eletrik dị n'ihu ụlọ ya.		
sammy	Ọ DỊ Paul mkpa bọlb ọkụ eletrik dị n'ihu ụlọ ya.		
	He also needed to clean the outside upstairs windows — his wife had mentioned this several times.	92.37	A
codedcheat	Ọ dịkwa ya mkpa ihicha windo ndị dị n'ụlọ elu site n'èzí — nwunye ya ekwuwo nke a ọtụtụ ugboro.		
lizzy	Ọ dịkwa ya mkpa ihicha windo ndị dị n'ụlọ elu n'èzí — ya ekwuwo nke a ọtụtụ ugboro.		
sammy	ya mkpa ihicha windo ndị dị n'ụlọ elu site n'èzí — nwunye ya ekwuwo nke a ọtụtụ ugboro.		

5. Conclusion

The development of a multilingual translation system for Nigerian languages represents a practical step toward creating essential NLP resources for under-resourced and endangered local languages. This study shows that combining automated tools like the similarity checker with tokenization, translation memory, and human evaluation produces accurate and reusable bilingual datasets. The test results, with similarity scores between 79% and 98%, confirm the system's reliability and its value for building parallel corpora. By supporting translators with clear workflows and quality checks, the system helps preserve Nigeria's linguistic diversity and reduces the risk of language extinction. It also lays a foundation for future applications such as machine translation and educational tools. Moving forward, further work should develop tailored evaluation metrics and explore crowdsourcing to involve native speakers and experts, ensuring continuous improvement and broader reach. The proposed system, therefore, is a promising contribution to safeguarding Nigeria's languages and expanding digital access for diverse communities.

References

Abdullahi, M. (2016). *A multilingual translation system for enhancing agricultural e-extension services delivery*. Semantic Scholar. Retrieved from https://api.semanticscholar.org/CorpusID:209503673

Afolabi, O. (2023). *Nigerian languages going extinct*. Free Knowledge Africa. Retrieved from https://freeknowledgeafrica.org/nigerian-languages-going-extinct/

Agantiem, A. (2017). Language (in) equality, language endangerment and the threats to Nigerian languages. *Journal of Literature, Languages and Linguistics*, 37, 21–28.

Ayegba, S. F., Musa, U., & Philip, N. (2016). Design and implementation of a system for automatic translation of Igala to English language. *International Journal of Scientific Research in Science and Technology*, 2(3), 212–216. https://ijsrst.com/paper/399.pdf https://doi.org/10.32628/ijsrst162325

Choudhary, N. (2021). LDC-IL: The Indian repository of resources for language technology. *Language Resources and Evaluation*, 55(3), 855–867. https://doi.org/10.1007/s10579-020-09523

Demas, H. (2016). *Making Amharic to English language translator for iOS*. Theseus.fi. Retrieved from https://www.theseus.fi/bitstream/10024/109553/1/Demas_Hana.pdf

Desai, N. P., & Dabhi, V. K. (2022). Resources and components for Gujarati NLP systems: A survey. *Artificial Intelligence Review*, 55(7), 1–19. https://doi.org/10.1007/s10462-022-10167-8

Eberhard, D. M., Gary, F., Simons, G. F., & Fennig, C. D. (2023). *Languages of the World, Twenty-sixth Edition*. Ethnologue. Retrieved from https://www.ethnologue.com

Fagbolu, O. O., Alese, B. K., Ogundele, A. O., & Adeyanju, K. A. (2015). *Digital Yoruba corpus*. ResearchGate. Retrieved from https://www.researchgate.net/publication/336274457_Digital_Yoruba_Corpus

Hou, Q., & Li-Hong, Z. (2022). Design and implementation of interactive English translation system in Internet of Things auxiliary information processing. *Wireless Communications and Mobile Computing*, 2022, 1–12. https://doi.org/10.1155/2022/3987970

Khurana, S., Koli, A., Khatter, K., & Singh, A. (2022). Natural language generation: State-of-the-art, trends, and challenges. *Journal of Information and Optimization Sciences*, 43(4), 1033–1057. [Check actual publication for exact source if needed]

Krishna, G. G. (2023). *Multilingual NLP (Term Paper)*. ResearchGate. Retrieved from https://www.researchgate.net/publication/369917117_Multilingual_NLP_Term_Paper

Olajoke, A. S., & Idowu, A. S. (2013). Translation and its linguistic implications for Yoruba/English bilinguals. *Learning*, 3(15), 99–104.