Assessing AI-Based Yorùbá Translation: A Computational and Human Evaluation Approach

Olayinka Modinat EDUN[1]& Festus Moses ONIPEDE[2]
[1,2]Department of General Studies, Federal Polytechnic Ilaro, Nigeria
***Corresponding email :**olayinka.edun@federapolyilaro.edu.ng

Introduction

The rise of artificial intelligence (AI)-driven machine translation has significantly advanced cross-linguistic communication. However, AI translation of Yorùbá—a tonal and morphologically complex language—remains challenging. This study evaluates AI-based Yorùbá translation using a combined computational and human assessment approach. Leveraging BLEU, METEOR, and TER for computational evaluation, alongside expert human judgments on fluency, accuracy, and cultural appropriateness, the study examines AI translations from Google Translate, Microsoft Translator, and NLLB-200. Artificial intelligence-driven machine translation (MT) has made significant strides in global linguistic accessibility. However, the translation of morphologically rich and tonal languages like Yorùbá presents unique challenges. Yorùbá's tonal system, noun-class structure, and agglutinative morphology often lead to errors in AI-generated translations, reducing their communicative effectiveness. Existing AI translation models, such as Google Translate, Microsoft Translator, and Meta's NLLB-200, are primarily trained on data from high-resource languages, leading to underrepresentation of Yorùbá's linguistic complexities. Thus, this study seeks to assess AI-based Yorùbá translation using a dual evaluation approach—computational metrics and human judgment—to determine the effectiveness and limitations of these models.AI-based MT has evolved from statistical approaches to neural networks and deep learning models, improving translation accuracy. Key challenges in Yorùbá AI translation include: tone representation (Yorùbá is a tonal language, and misplacement of tone marks can alter meaning (e.g., ọ̀ràn "case" vs. ọ̀ràn "problem")), morphological complexity (Agglutination and compounding in Yorùbá pose difficulties for AI segmentation and parsing), and idiomatic expressions (Yorùbá proverbs and idiomatic phrases often lose meaning in AI translations due to lack of cultural context).

Materials and Methods/Methodology

Translation samples were drawn from Yoruba proverbs collections (e.g., Owomoyela 2005), Yoruba language textbooks (e.g., Bámgbóṣé 1966, Awobuluyi 2008), and Nigerianprimary school readers. Three AI translation models—Google Translate, Microsoft Translator, and NLLB-200—were used to translate these texts into English and vice versa. There were three commonly used metrics utilized: The Bilingual Evaluation Understudy (BLEU) calculates the amount of word overlap between reference and AI-generated translations; Word order, stemming, and synonyms are taken into account by METEOR (Metric for Evaluation of Translation with Explicit Ordering), whereas TER (Translation Edit Rate) determines how many adjustments are necessary to match the reference translation. Human evaluation was used. A group of linguists and Yorùbá language specialists evaluated AI translations according to three criteria: cultural appropriateness (maintaining idiomatic expressions and tonal accuracy), accuracy (faithfulness to source meaning), and fluency (grammar and syntactic correctness).

Results and discussion

Example notes from evaluators from *Google Translate* omitted tonal marks, leading to wrong interpretations. *Microsoft* handled basic syntax better but struggled with idioms. *NLLB-200* preserved more Yoruba-specific constructions and tone-sensitive expressions. Common errors included tone misrepresentation where AI translations frequently omitted diacritical marks, leading to altered meanings. Word-for-word translation often failed to capture contextual meaning and idiomatic expression in Yorùbá proverbs and fixed expressions were inadequately translated. The BLEU, METEOR, and TER scores for the three AI models as presented in Table 1 make NLLB-200 outperformed the other models, particularly in METEOR and TER, suggesting better word alignment and lower edit distance.

Conclusion

This study assessed AI-based Yorùbá translation through computational and human evaluation. While AI models demonstrated moderate proficiency, significant gaps persist in tonal representation, idiomatic translation, and contextual accuracy. Results reveal gaps in tonal fidelity, idiomatic expressions, and morphological agreement, highlighting the need for improved AI models. The findings contribute to computational linguistics, translation studies, and the development of more effective Yorùbá translation systems.