Ẹhugbo Ka! Advancing Machine Translation for the Low-Resource Ẹhugbo Language through Parallel Corpus Development

Ukachi Agnes Eze-Mbey UAE@ANDREW.CMU.EDU
Carnegie Mellon University Africa

Uloma Calista Eze-Mbey UCEZE-MBEY@STUDENT.OAUIFE.EDU.NG
Obafemi Awolowo University

Ololade Anjuwon OLOLADE@DATASCIENCENIGERIA.AI
Data Science Nigeria

Abstract

Despite advancements in language technologies, there consistently seems to be an exclusion of low-resource African languages and their dialects like Ẹhugbo, a critically endangered variant of Igbo spoken by fewer than 150,000 people in Afikpo, Nigeria. This exclusion perpetuates social and linguistic inequities, leaving speakers of such dialects without access to digital tools that could preserve their language and culture. This paper presents Ẹhugbo Ka! ("Greetings Ẹhugbo!") addresses this gap. We gathered and built the only publicly available parallel corpus, 1,021 Ẹhugbo-English sentences from the New Testament of the Bible, we evaluated and fine-tuned two state-of-the-art models, M2M100 (facebook/m2m100 418M) and NLLB (facebook/nllb-200-distilled-600M). Initial results were stark: M2M100 achieved a BLEU score of 1.2188, while NLLB scored only 0.0262. After fine-tuning, M2M100 improved to 16.1719, and NLLB achieved 20.4016, demonstrating the potential of adapting LLMs for low-resource languages. Our findings reveal both promise and challenges. While fine-tuning significantly improves performance, the lack of diverse datasets limits translation quality and reinforces the need for inclusive data collection practices. This work highlights the importance of community-driven approaches, as linguistic preservation cannot be achieved without the active involvement of native speakers.This project not only advances the field of low resource MT but also serves as a call to action for researchers and developers to prioritize linguistic diversity, ensuring that no language is left behind in the digital age.

Keywords: multilingual low resource, resources for less-resourced languages, minoritized languages, less resourced languages, endangered languages, indigenous languages, corpus creation, multilingual corpora, evaluation, datasets for low resource languages, Igbo, Igbo language.

1. Introduction

There are 7,164 languages worldwide Ethnologue (2025), of which about 3,000 are African. Nigeria alone has about 525 languages, with Igbo language having over 31 million speakers Wikipedia (2025). Towards the end of 2006, the United Nations predicted that some minor languages of the world would go extinct in the next 50 years. On this list was the Igbo language spoken in southeastern Nigeria by over 20 million people. The pervasive issue of Igbo language being relegated to secondary status raised concerns to the extent that

UNESCO has projected a risk of Igbo language becoming extinct by 2025 Asonye (2013). A major factor contributing to this is the multi-dialectal nature of the language Nwaozuzu (2008), which has made it challenging for linguistic initiatives, lexical tools, and language technologies that solely focus on the 'Standard Igbo' to gain widespread acceptance, particularly among the broader language speaking community. Dialects, often overshadowed by dominant languages, carry the unique cultural identities, histories, and worldviews of their speakers. They are vital to the diversity of human expression and the richness of global heritage. Yet, in the digital age, many dialects just like Ẹhugbo face marginalization, exclusion, and even extinction, as they are of ten overlooked in favor of standardized languages. This exclusion perpetuates inequalities, denying speakers of dialects access to information, education, and opportunities in the digital space. Recognizing and preserving dialects is therefore essential to ensuring linguistic equity, cultural sustainability, and the full participation of all communities in the digital revolution.
The Ẹhugbo dialect also known as Afikpo is a dialect of Igbo language spoken in Afikpo North Local Government Area of Ebonyi State, South-East Nigeria. Ẹhugbo is distinct in its linguistic features, including two additional alphabets beyond Igbo's 36, resulting in a richer and more complex lexicon. Spoken by fewer than 150,000 people in Ebonyi State, it is critically endangered. While it is the first language of older generations, younger speakers are increasingly shifting to English and Nigerian Pidgin according to Orife (2020), leading to a decline in its use and transmission. With limited access to digital resources and essential information, Ẹhugbo speakers risk cultural isolation and exclusion from the benefits of the digital age. Motivated by a commitment to linguistic inclusion and cultural equity, this research aims to empower Ẹhugbo speakers by creating the first Ẹhugbo-English parallel corpus. This corpus will serve as a foundational resource for training a robust MT system. By harnessing the power of generative AI models adapted to African languages, ensuring that the Afikpo community can access the wealth of information available in the digital world.

2. Background and Related Work

Igbo language is often categorized among the "left-behind" languagesJoshi et al. (2020), indicating its limited representation in language technologies and the scarcity of available datasets, though efforts to develop lexical resources have been ongoing, with contributions from scholars Ogbalu (1962), Nnaji (1985), Igboanusi (2017) and Mbah (2021). In the realm of natural language processing (NLP), significant strides have been made with modern day datasets Onyenwe et al. (2018), Ezeani et al. (2020), Adelani et al. (2022). Early foundational works, such as the early dictionaries Ogbalu (1962), Nnaji (1985) Eke (2001). laid the groundwork for later advancements. More recently, a benchmark dataset containing 5,630 English sentences translated into Igbo was developed Ezeani et al. (2020), alongside 5,503 Igbo sentences translated into English, forming a bilingual corpus. The JW300 dataset Agić and Vulić (2019) also contributed a large scale corpus, primarily focused on religious texts. Also, the IgboSum1500 project by Mbonu et al. (2022) introduced a dataset for text summarization, comprising 1,500 articles. A groundbreaking development is the IgboAPI project Emezue et al. (2024), which incorporates various Igbo dialects, a significant advancement from previous works that predominantly focused on Standard Igbo. This dataset includes

1,066 words specifically from the Ẹhugbo dialect. Despite these advancements, the role of dialectal diversity in shaping language technologies, such as lexicons or MT systems, remains underexplored within the Igbo context. Research in other languages, however, offers valuable insights. For example, multi-dialectal neural machine translation for Japanese dialects were investigated Abe et al. (2018), demonstrating its potential to benefit populations more familiar with regional variations. Similarly, translation challenges between Egyptian Arabic and Modern Standard Arabic have been addressed Almansor and Al-Ani (2017). Building on these, our work focuses on the Ẹhugbo dialect of Igbo, leveraging a dialectal aware dataset to conduct experiments in Igbo-English translation. By fine-tuning pre-trained multilingual models such as M2M100 and NLLB, we explore the feasibility of developing a robust translation system for Ẹhugbo, a language variant that has historically been excluded from mainstream NLP efforts. Our study contributes to the growing body of work on Igbo language technology while highlighting the importance of incorporating dialectal diversity into MT systems to ensure fairness and inclusivity for underrepresented linguistic communities

3. Methodology

3.1. Pre-training and Dataset

We translated 1021 sentences [1] from the Ẹhugbo New Testament Bible (Baibulu Nso Agba Ohuu Na Okwu Ẹhugbo) from the books of Matthew, Philemon, Jude, 1st-3rd John, and Ephesians.Though the Bible was written with about 7 translations as stated in its preface, we translated it using the most context-preserving version which was the New International Version(NIV). A sample can be seen in Table 1. We have adopted a transfer learning approach by fine-tuning pretrained M2M100 (facebook/m2m100 418M) and NLLB (facebook/nllb-200-distilled-600M) models, harnessing their rich intrinsic knowledge to improve Ẹhugbo translation performance. These sentences were divided into 80% of the training set, and 10% each for the Dev and Test sets.

3.2. Model Training

We utilize the Hugging Face Transformers library for its enhanced capabilities and seamless model integration with state-of-the-art Seq2Seq modeling. The training process has been carefully designed to optimize model performance, using the following key steps:

• **Hyper-parameter tuning**: We perform experiments with different hyper-parameters like the number of epochs, batch sizes, and learning rate, within the limits of the data to identify an appropriate configuration for the Ẹhugbo translation.

• **Regularization**: Considering the size of the dataset, we implemented a dropout regularization technique to avoid over-fitting and improve model generalization.

• **Loss function**: An appropriate loss function, the cross-entropy with label smoothing, is selected to guide model learning to produce translations.

• **Evaluation and refinement**: Realizing the limitations of using a small dataset for

1. Access the `English-Ehugbo Sentences` on https://huggingface.co/datasets/Ukachi/NLP-Ehugbo/resolve/main/English-Ehugbo.csv

evaluation, we monitor the performance of the model throughout training using established metric such as the BLEU score(bilingual evaluation understudy), an algorithm for the evaluation of the quality of text that has been machine-translated from one natural language to another, and human evaluation to assess translation quality, fluency, and cultural appropriateness. Our methodology was to create a functional Ẹhugbo-English MT system by fine-tuning pre-trained M2M100 and NLLB models with a limited dataset. Despite the amount of data, our system has shown promising results. These indicate its ability to produce translations.

Table 1: Sample of Translated Dataset

S/N	Ẹhugbo	English
1	Ana m ederi unu ihie na ohu bayeri ndem na-achọ na wo tilu unu ihu.	I am writing these things to you about those who are trying to lead you astray.
2	Ngọzi dịịrị ndem nọ na erim-ụjụ, maka na a je-akasi wo obu.	Blessed are those who mourn, for they will be comforted.
3	Ayị na-ahụtarị onwaayị na ẹnya maka na Chineke bururi ụzọ hụ ayị na ẹnya.	We love because he first loved us.
4	Ya o je je-ekute Meri kulete na ulo e, lụrụ a ya ọ bụrụ nyee ye.	He took Mary home as his wife.

4. Results and Discussion

To evaluate the performance of translation models, both the M2M200 and NLLB models were tested on our dataset of 1,021 sentences. The M2M200 model achieved a BLEU score of 1.2189, while the NLLB model yielded a significantly lower score of 0.1313. These results highlight the disparity in translation effectiveness between the two models and emphasize the need for further optimization to enhance their accuracy and reliability. The models were then fine-tuned using distributed training with a single GPU. For 10 epochs each, the training process of the M2M100 model took approximately 16 minutes and 2 seconds while that of the NLLB model took 20 minutes 59 seconds. The training loss for the M2M100 model decreased from 2.9523 to 0.3474, while that of theNLLB model decreased from 3.0180 to 1.03570. The evaluation and prediction metrics of the models after 10 epochs are as seen in Table 2 and Table 3.

4.1. Discussion of Results

The results of our study highlight both the feasibility and the challenges of fine-tuning pre-trained multilingual models, specifically M2M100 and NLLB, for the MT of Ẹhugbo-English with a limited dataset. Despite the constrained data, the M2M100 and NLLB models achieved promising results, as demonstrated by evaluation BLEU scores of 19.3679 and 20.4016 respectively and predictive BLEU scores of 16.015 and 15.3053 respectively.

These scores suggest that with additional data, the model could improve its ability to generate translations that align more closely with human output.
However, our findings also underscore key fairness concerns in Large Language Models (LLMs) when handling low-resource and dialectal languages. The relatively low BLEU scores raise critical questions about biases in multilingual models, particularly in their ability to equitably represent and translate linguistically diverse communities. Several factors contribute to this challenge. The small dataset size limited the model's capacity to fully learn the nuances of Ẹhugbo, adialect of Igbo, despite their lexical similarities.
Our study serves as an important step toward building more inclusive and fair MT systems. Byfine-tuning pre-trained models and expanding our dataset, we aim to improve both translation accuracy and cultural sensitivity. More broadly, our work highlights the necessity of fair representation for underserved languages in LLMs, ensuring that technological advancements do not disproportionately benefit widely spoken languages while neglecting marginalized linguistic communities. Addressing these fairness gaps is crucial for developing equitable AI systems that serve diverse populations. The evaluation and prediction metrics of the models after 10 epochs are as seen in Table 2 and Table 3.

Table 2: Evaluation Metrics Results

Metric	***M2M100***	***NLLB***
Evaluation BLEU Score	19.3679	20.4016
Evaluation Generation Length	37.7282	37.9806
Evaluation Loss	2.5129	2.5186
Evaluation Runtime	0:00:44.11	0:00:41.20
Evaluation Samples	103	103
Evaluation Steps Per Second	0.589	0.631

Table 3: Prediction Metrics Results

Metric	***M2M100***	***NLLB***
Prediction BLEU Score	16.015	15.3053
Prediction Generation Length	42.5588	42.3039
Prediction Loss	2.7003	2.6996
Prediction Runtime	0:00:46.95	0:00:44.22
Prediction Samples	102	102
Prediction Samples Per Second	2.173	2.307
Prediction Steps Per Second	0.554	0.588

4.2. Key Findings

- **Performance Evaluation**: The developed parallel corpus demonstrated strong translation capabilities, as reflected in its BLEU scores. The M2M100 model achieved an evaluation BLEU score of 16.1719, indicating its accuracy, and a prediction BLEU

score of 11.8838. In comparison, the NLLB model outperformed it, attaining an evaluation BLEU score of 20.4016 and a prediction BLEU score of 15.3053, underscoring its higher reliability.

- **Translation Quality**: Despite the constraints of a limited dataset, our translations exhibit fluency and clarity, demonstrating the robustness and effectiveness of our approach.
- **Cultural Sensitivity**: Our parallel corpus prioritizes the preservation of cultural nuances, ensuring that translated content maintains linguistic authenticity. To further enhance this, we plan to expand our training data to 10,000 parallel sentences and actively engage local stakeholders. This will enable the system to better address the unique linguistic and cultural needs of the Ẹhugbo-speaking community, fostering a coordinated and inclusive effort toward high-quality MT.
- **Social Impact**: Beyond technical advancements, our research aims to empower Ẹhugbo speakers, enabling their full participation in the global digital economy while safeguarding their linguistic heritage. This initiative aligns with the broader goal of advancing the digital presence of African languages, ensuring their sustainability in an increasingly AI driven world.

4.3. Limitations

- **Unique Alphabets in Ẹhugbo**: Ẹhugbo, though a dialect of Igbo, includes two additional alphabets Ẹ and Ụ which are not present in standard Igbo. These unique characters are often treated as foreign text by the model during testing, potentially affecting translation accuracy and performance.
- **Limited Training Data:** This work was trained exclusively on religious data, specifically sentences from the New Testament of the Bible. This limitation arises from the scarcity of standardized texts in Ẹhugbo, as only four books are currently written strictly in the dialect. The lack of diverse textual sources restricts the model's ability to generalize across different domains.
- **Bible Translation Variability:** Though the initial body of work was based on multiple versions of the bible (e.g., NIV, NKJV, KJV, NLT, GNBUK), each with slight variations in wording and phrasing, the parallel English sentences were based on the 2011 NIV Bible. While the overall context is preserved, precise word choices may differ, introducing inconsistencies into the training data.
- **Model Training on Diverse Data:** The models used in this work (m2m100 and NLLB) were pre-trained on more diverse datasets, such as news data, which may not align well with the religious and cultural context of the Ẹhugbo Bible. This mismatch could limit the models' effectiveness in capturing the nuances of Ẹhugbo.
- **Limited Model Testing**: Only two models were tested in this study. While the results are promising, evaluating additional models could provide a more comprehensive understanding of their performance on Ẹhugbo translation tasks.

- **Small Dataset Size:** The dataset used for this work consists of slightly over 1,000 sentences. While this is a significant starting point, there is considerable room for expansion. A larger and more diverse dataset would likely improve the model's accuracy and robustness.

5. Conclusion

This research highlights the critical role of linguistic resources such as corpora, terminologies, and dictionaries in enhancing MT for low-resource languages. By fine-tuning pre-trained M2M100 and NLLB models with a limited dataset, we demonstrate the feasibility of adapting large-scale models to underrepresented languages. The M2M100 model achieved an evaluation BLEU score of 16.1719 and a prediction BLEU score of 11.8838, while the NLLB model outperformed it with an evaluation BLEU score of 20.4016 and a prediction BLEU score of 15.3053. These results underscore the effectiveness of leveraging pre-trained models, even in resource-scarce scenarios.

However, our findings emphasize the necessity of high-quality linguistic resources. Expanding the dataset to include 10,000–15,000 parallel sentences from diverse domains—beyond religious texts—would significantly enhance model performance. Future work will focus on curating domain specific corpora, refining terminologies, and integrating bilingual dictionaries to improve translation accuracy and contextual fidelity. Additionally, investigating domain adaptation techniques and alignment strategies will further optimize MT systems for low-resource settings.

Ultimately, this research contributes to the broader effort of developing equitable and culturally inclusive MT solutions. By strengthening linguistic resources and advancing MT methodologies, we aim to support linguistic diversity, foster accessibility, and enable more effective cross lingual communication for underrepresented languages.

6. Sustainability Statement

In alignment with the principles of Green AI Schwartz et al. (2019), our work prioritizes computational efficiency and environmental responsibility in MT model training. To minimize the environmental impact, we employed distributed training on a single GPU, significantly reducing energy consumption while maintaining model performance.

The fine-tuning process was optimized for efficiency, with M2M100 training completing in 16 minutes and 2 seconds and NLLB in 20 minutes and 59 seconds, across 10 epochs. Our approach ensures a balance between model accuracy and sustainability.

We provide a carbon impact assessment Henderson et al. (2020) to promote transparency in computational costs. The training process utilized a TPU v4 with an estimated energy consumption of 0.0977 kWh and a total carbon footprint of 87.99 gCOe, considering a power usage effectiveness (PUE) factor of 1.12 in the cloud infrastructure (GCP, South Africa region).

By adhering to these sustainable AI practices, we contribute to lowering the environmental cost of Machine Translation research while advocating for energy-efficient and resource-conscious approaches in NLP model development. Future work will explore further reductions in computational overhead through optimized architectures.

Acknowledgments

We gratefully acknowledge the Nigeria Bible Translation Trust (NBTT), the Ehugbo Literacy and Bible Translation Project Committee, and the Initiative on Mother Tongue and Literacy Development for their invaluable contribution to this research through the Ehugbo New Testament Bible—the foundational corpus of this work. Their dedication to preserving and promoting the Ehugbo language has made this study possible.
We extend special appreciation to Pastor Victor Elem for his mentorship, linguistic insights, and unwavering support throughout this research. His expertise in Ehugbo language documentation and Bible translation profoundly shaped the direction of this work.
Finally, we acknowledge the welcoming support from the Afikpo community. This research would not have been possible without their commitment to indigenous language preservation.

References

K. Abe et al. Multi-dialect neural machine translation and dialectometry. In *Proc. 22nd Conf. Comput. Nat. Lang. Learn.*, pages 93–104, 2018.

D. I. Adelani et al. A few thousand translations go a long way! leveraging pre-trained models for african news translation. *arXiv*, arXiv:2205.02022, 2022.

Ž. Agić and I. Vulić. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, pages 3204–3210, 2019.

E. H. Almansor and A. Al-Ani. Translating dialectal arabic as low resource language using word embedding. In *Proc. Int. Conf. Recent Adv. Nat. Lang. Process.*, pages 32–38, 2017.

E. Asonye. Unesco prediction of the igbo language death: Facts and fables. *J. Afr. Lang. Stud.*, 12(3):45–62, 2013.

J. Eke. Igbo-english dictionary, 2001. URL http://www.igbodictionary.com. Accessed: Jan. 20, 2025.

C. C. Emezue et al. The igboAPI dataset: Empowering igbo language technologies through multi-dialectal enrichment. *arXiv*, arXiv:2405.00997, 2024.

Ethnologue. Languages of the world, 2025. URL https://www.ethnologue.com. Accessed: Jan. 20, 2025.

I. Ezeani et al. Igbo-english machine translation: An evaluation benchmark. *arXiv*, arXiv:2004.00648, 2020.

P. Henderson et al. Towards the systematic reporting of the energy and carbon footprints of machine learning. *J. Mach. Learn. Res.*, 21(248):1–43, 2020.

H. Igboanusi. *English-Igbo Glossary of HIV, AIDS and Ebola-related Terms*. Kraft Books, Ibadan, Nigeria, 2017.

P. Joshi et al. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, pages 6282–6293, 2020.

B. M. Mbah. *Igbo Dictionary Osanye Okwu Igbo na Nkowa Ya*. Cidjap Press, Enugu, Nigeria, 2021.

C. Mbonu et al. Igbosum1500: Introducing the igbo text summarization dataset. In *3rd Workshop Afr. Nat. Lang. Process.*, 2022.

H. I. Nnaji. *Modern English-Igbo Dictionary*. Longman Nigeria, Lagos, Nigeria, 1985.

G. I. Nwaozuzu. *Dialects of the Igbo Language*. University of Nigeria Press, Nsukka, Nigeria, 2008.

F. C. Ogbalu. *Kwa-Okwu: Igbo English-English Igbo Dictionary*. University Publishing Co., Onitsha, Nigeria, 1962.

I. E. Onyenwe et al. A basic language resource kit implementation for the igboNLP project. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 17(2), 2018.

I. Orife. Towards neural machine translation for edoid languages. *arXiv*, arXiv:2003.10704, 2020.

R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni. Green AI. *arXiv*, arXiv:1907.10597, 2019.

Wikipedia. Igbo language, 2025. URL https://en.wikipedia.org/wiki/Igbo_language. Accessed: Jan. 21, 2025.