# University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR)

**ISSN: 2714-3627**

***A Journal of the Faculty of Computing, University of Ibadan, Ibadan, Nigeria***

**Volume 16 No. 1, January 2026**

**journals.ui.edu.ng/uijslictr**
**http://uijslictr.org.ng/**
**uijslictr@gmail.com**

**University of Ibadan**
**Journal of Science and Logics in ICT Research**

# An Enhancement of African Low-Resource Corpora with NLP IgboT5

**Jacinta Chioma Odirichukwu[1,*], Reginald Nnadozie Nnamdi[2], Simon Peter Chimaobi Odirichukwu[3]**

[1] *Department of Computer Science, Federal University of Technology, Owerri, Imo State, Nigeria*
[2] *Department of Philosophy, Veritas University, Abuja, FCT, Nigeria*
[3] *Department of Health,Primary Health Development Agency, Owerri, Imo State, Nigeria*
[*]*Correspondence: jacinta.odirichukwu@futo.edu.ng*

***Abstract***
This paper adopts the Text-to-Text Transfer Transformer (T5) for the Igbo language Natural Language Processing Tasks. IgboT5 enhances the previous digital Igbo Thesaurus through the creation of a high-quality Igbo dataset. The paper fine-tunes a multilingual T5 model and evaluates it on tasks such as definition generation, paraphrasing, translation, and context completion. This paper contributes to the advancement of low-resource African languages and opens doors for future Natural Language Processing (NLP) applications.

***Keywords***: T5, Igbo Dataset, NLP, Transformer Model, IgboT5

## 1. Introduction

The Igbo language is one of the major native languages in Nigeria. It has a well-established lexical and grammatical structure that is profoundly interwoven with Igbo cultural identity. The availability of digital resources that explain the meaning of Igbo words beyond mere translation is very limited. Previously, the Igbo Thesaurus introduced a web based monolingual dictionary, which provides definitions and synonyms of Igbo words, rather than English translations Odirichukwu & Nnamdi [1].

Natural Language Processing (NLP) involves developing algorithms that enable computers to understand, process, and interpret natural languages. It has many applications such as text processing, sentiment analysis, named entity recognition (NER), machine translation, speech recognition, text-to-speech conversion, and others. Text summarization, virtual assistants, chatbots, and information retrieval are among the key NLP applications Sawicki et al.[2]. Odirichukwu et al.[11] reviews the existing literature works on Igbo NER, highlighting the challenges, creating opportunities and looking into the potential applications of NER in developing Igbo digital assistants, intelligent search, and machine translation.

Building upon our previous work, this paper creates and analyzes a robust Igbo dataset. The Text-to-Text Transfer Transformer will be used in future work to model and generate Igbo text for various NLP applications. This framework will enable building models for various Igbo text-based tasks through extensive exploratory data analysis, preprocessing, tokenization, and various natural language processing applications. With the creation of a high-quality Igbo dataset and application of state-of-the-art NLP methods, IgboT5 aims to enhance the performance of Igbo language models and pave the way for future research.

The following are the objectives of this research:

1) To improve the existing Igbo Thesaurus by creating a high-quality Igbo dataset.

Jacinta Chioma Odirichukwu, Reginald Nnadozie Nnamdi, and Simon Peter Chimaobi Odirichukwu (2026). An Enhancement of African Low-Resource Corpora with NLP IgboT5, *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR),* Vol. 16 No. 1, pp. 47 - 54

2) To preprocess the created dataset by fine-tuning it to be suitable for training T5 models for linguistic diversity and Igbo domain coverage.
3) To tokenize Igbo words to handle Igbo text-to-text generation.
4) To pave the way for future research on Igbo NLP applications.

This paper is organized as follows: Section II presents materials and methods. Section III discusses the results. Section V concludes the study.

## 2. Related Works

In the classification of Natural Language Processing, there are two main categories: high-resource languages and low resource languages. Among these groupings, African languages are categorized as low-resource languages. The major reason for this categorization is attributed to limited investments, which have resulted in little interest from researchers in these languages [3].

IgboBERT is a model acclaimed to be the first and only transformer-based language model pre-trained on the Igbo language. It was fine-tuned using a downstream NER task with the MasakhaNER dataset. IgboBERT achieved significant performance improvements when compared with other parameters, having been trained with 84M parameters. Comparing other models such as MEET, XML-R, and DistilBERT with IgboBERT on pre-training with relatively small raw data, the results indicate that there is no convergence in training against validation loss. With the result of about 77.94M achieved by IgboBERT with small raw data, the introduction of more data for fine-tuning may likely improve performance further [4].

Text-to-Text Transfer Transformer (T5) has been operational through a unified text-to-text format and scale, which helps achieve state-of-the-art results on a variety of English language NLP tasks. To enhance results, mT5, a multilingual variant of T5, was introduced and pre-trained on a new Common Crawl-based dataset covering over 101 languages. To achieve the best results, mT5 was detailed and modified to produce expected results. The results showed high performance on many multilingual benchmarks. The setting of mT5 was made simple to prevent accidental translation in zero shot settings, where generative models might partially translate predictions into the wrong language [5].

Multilingual T5 (mT5), which pre-trains a sequence-to sequence model on massive monolingual texts, shows promising results on many cross-lingual tasks and can be improved by introducing multilingual text-to-text transfer Transformer with translation pairs (mT6). mT6 explored three cross-lingual text to-text pre-training tasks: machine translation, translation pair span corruption, and translation span corruption. There was also further evaluation and proposal for non-autoregressive objectives in text-to-text pre-training. Furthermore, eight multilingual benchmark datasets for sentence classification, named entity recognition, question answering, and abstractive summarization were evaluated. Results indicate significant improvement in mT6's cross-lingual transferability over mT5 [6].

In the transfer learning approach, the unified transformer framework (T5) has been tested in converting all language problems to a text-to-text format. A multilingual version of the T5 model (mT5) was also introduced. However, its effectiveness has not been well developed for non-English tasks with diverse data. To confirm the effect on non-English tasks, mT5 was applied to a wide range of languages with different dialects, such as Arabic. The introduction of Arabic led to a novel benchmark for Arabic language generation (ARGEN). ARGEN covered seven important tasks. To compare models, three important Arabic T5-style models were pre-trained and evaluated with ARGEN. Results showed that with 49% less data, the new models performed significantly better than mT5 on all ARGEN

tasks (in 52 out of 59 test sets) and set several new state-of-the-art results [7].

In Natural Language Processing, Transformer Language Models (TLMs) have become an integral part of systems. Despite the fact that many Transformer models have been introduced to serve many languages, the need for efficient models persists due to the shortage or lack of models for pre-training low-resource or indigenous languages. Seeking solutions, IndT5, the first Transformer language model for indigenous languages, was introduced. To train IndT5, IndCorpus, a new dataset for 10,000 indigenous languages and Spanish, was built. IndT5 was applied to machine translation by investigating numerous approaches to translate between Spanish and indigenous languages. Results showed that IndT5 and the IndCorpus dataset for 10,000 indigenous languages and Spanish, submitted to the AmericasNLP 2021 shared task, significantly improved translation. The only constraint was the absence of parallel and monolingual data [8].

A Recurrent Neural Network (RNN)-based Neural Machine Translation (NMT) system was introduced for English-to-Igbo translation to address numerous issues in low-resource language tasks. One major highlight is the effectiveness of Long Short-Term Memory (LSTM) models combined with attention mechanisms, specifically dot-product attention. Dot-product attention plays a vital role in producing fluent and semantically accurate translations. Though trained on limited data, the model achieved good performance comparable to Tatoeba and JW300. Extending the research to English-French datasets, the RNN surpassed the Tatoeba English benchmark by 9 BLEU points. Cross-lingual records underscore the robustness of applications across languages with varying grammatical complexity. The MarianNMT framework was introduced to enhance translation quality, yielding a BLEU score of 0.43, showing significant improvement of 4.83 points over existing HuggingFace English-Igbo baselines (Tatoeba and JW300), with approximately 70% semantic translation accuracy on an evaluation set of 597 samples. Results emphasized the potential of pre-trained transformer models in augmenting low resource NMT systems. However, future work should consider larger vocabulary datasets and multilingual NMT for many low-resource languages using alternative techniques [9].

The urgency of building models and datasets that can support text generation, translation, summarization, and other tasks in the Igbo language cannot be overemphasized. The need arises as text-to-text transfer (T5), which has been successfully applied to many natural language processing tasks, has not been extensively applied to Igbo. There are limitations in applying decoder-encoder architectures to the Igbo language due to low-quality Igbo corpora. Some cross-lingual datasets such as WikiLingua, XQuAD, TyDiQA, and others contain limited Igbo language data [10].

## 3. METHODOLOGY

IgboT5 is developed in three phases:

- Dataset Creation
- Preprocessing
- Task Framing

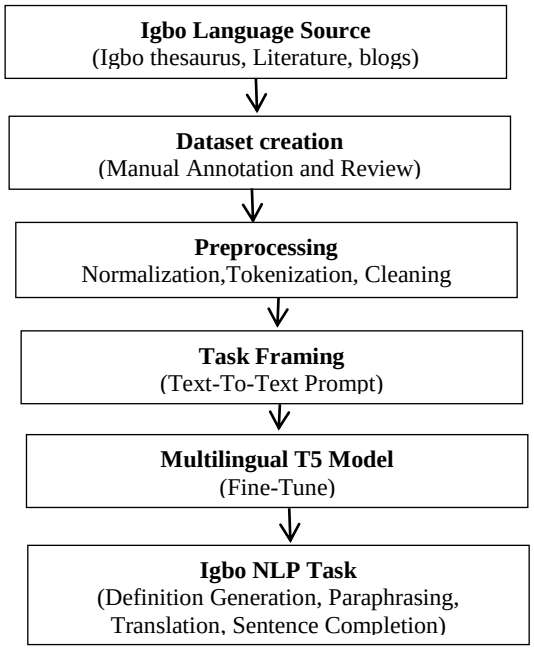


**Fig. 1: System Framework**

### *A. Dataset Creation*

The dataset of the Igbo corpus for this work was generated by Rev. Fr. Dr. Reginald Nnadozie Nnamdi, using his linguistic knowledge and improved by a team of Igbo scholars from research done through the Igbo Thesaurus, native literature, and online Igbo blogs. The IgboT5 dataset was manually annotated and reviewed by linguists. Igbo scholars ensured cultural and linguistic accuracy. This dataset was created and grouped for NLP tasks.

The dataset entries include:

• okwu (word): The Igbo word entry

• ngalaba (POS): The Igbo parts of speech

• mpta (meaning): The definition and usage examples

• cleaned tokens: Tokenized and processed version of the text

### *B. Preprocessing*

The raw dataset underwent the following preprocessing steps:

1) Normalization: Standardization of Igbo diacritics and tone marks took place to handle orthographic variants.
2) Tokenization: Training of the Igbo corpus subword tokenizer took place to create vocabulary for segmenting text in this low-resource language.
3) Cleaning: The dataset was cleaned to retain high-quality corpora by removing duplicates, distorted words and sentences, and other noise.
4) Task Framing: The corpus was framed into T5 schema, where each text was represented as an NLP prompt.

TABLE I

**EXCERPT FROM THE IGBO LEXICAL DATASET (4,052 ENTRIES)**

| ID | Okwu | ngalaba | mpüta | Cleaned Tokens |
|---|---|---|---|---|
| 0 | Abalị | (n) | Uchichi, anyasi, itiri. Ọchịchịrị | [’uchichi’, ’anyasi’, ’itiri’, ’ọchịchịrị’] |
| 1 | Abịdịị | (n) | Mkpụrụ edemede Alphabet: a,b,ch,d. | [’mkpụrụ’, ’edemede’,'Alphabet:a,b,ch,d.'] ] |
| 2 | Abọ | (n) | Akpati, Igbe ejiri ekete, kwee | [['Akpati', 'Igbe', 'ejiri', 'ekete', 'kwee', ]] |
| 3 | Abrakadabra | (n) | Magik. | [’magik’] |
| 4 | Absolut | (nkw) | Enweghị ngbagha, Akwaa akwụrụ, Ọkakaa. | [’Enwegh’, ’ngbagha’, ’Akwaa’, ’akwụrụ’, ’ọkakaa'] |
| . . . | . . . | . . . | . . . | . . . |
| 4047 | Ịtụgharị | (ngw) | Igbanwe, Ịnọgharị, Icheghari, Ịsụgharị | [’Igbanwe’, ’Ịnọgharị’, ’Icheghari’, ’Ịsụgharị’] |
| 4048 | AI | (n) | Ak Iheakambere. Akdre`e... | [’abbrev’, ’ak’...] |
| 4049 | Tụma | (nkw) | Kama. dr n’ite, ya dr... | [’kama’, ...] |
| 4050 | Ra | (ngw) | r, ra. Gbag (ntimiwu)... | [’nsianangw’, ’r’...] |
| 4051 | Obeji | (n) | Ngwr e ji awa nk. Onyike... | [’ngwr’, ’e’, ’ji’...] |

## 4. RESULTS AND DISCUSSION

### *A. Dataset Characteristics and Analysis*

The development of IgboT5 resulted in a comprehensive Igbo lexical dataset comprising 4,052 manually curated entries. Figure 1 illustrates the distribution of Igbo parts-of-speech (ngalaba) tags across the dataset, revealing a balanced representation of various grammatical categories. The dataset demonstrates a predominance of nouns (n), which aligns with the lexical nature of dictionary entries, followed by verbs (ngw) and other word classes (nkw). This distribution reflects the natural linguistic patterns in Igbo vocabulary and ensures comprehensive coverage for downstream NLP tasks.

The token length distribution analysis presented in Figure 2 provides crucial insights into the complexity and richness of the Igbo definitions. The histogram reveals that most lexical entries contain between 5-15 tokens, indicating substantial semantic content while maintaining manageable input lengths for T5 processing. This distribution is particularly favorable for transformer-based models, as it falls within optimal sequence lengths that balance computational efficiency and linguistic richness.

### *B. Task Performance and Model Capabilities*

The IgboT5 model demonstrates remarkable versatility across multiple NLP tasks, as illustrated in Tables II and III. The model's ability to generate contextually appropriate definitions, paraphrases, translations, and sentence completions represents a significant advancement in Igbo language processing capabilities. The definition generation task showcases the model's understanding of semantic relationships and cultural context. For instance, the definition of "Uchichi" accurately captures both temporal and cultural nuances associated with early morning in Igbo culture. The paraphrasing capabilities demonstrate sophisticated linguistic variation, employing different lexical choices while preserving semantic meaning. This is particularly significant for Igbo, where synonymous expressions often carry subtle cultural connotations.

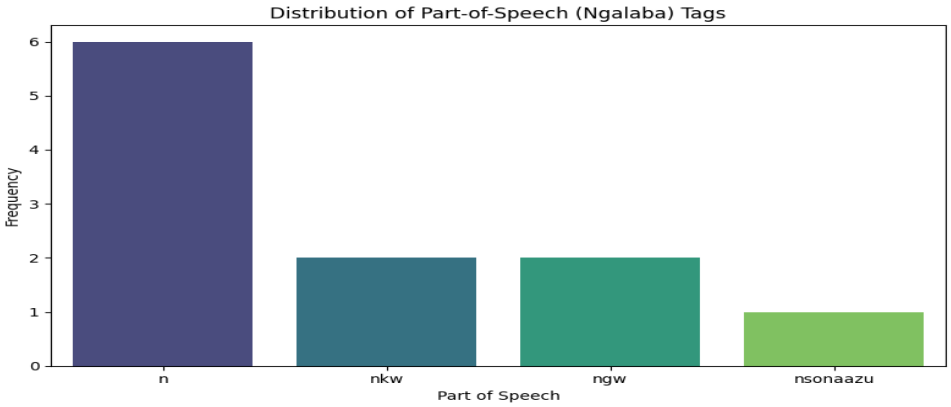


Fig. 2. Distribution of Igbo Parts-of-Speech (ngalaba) tags in Igbo Dataset

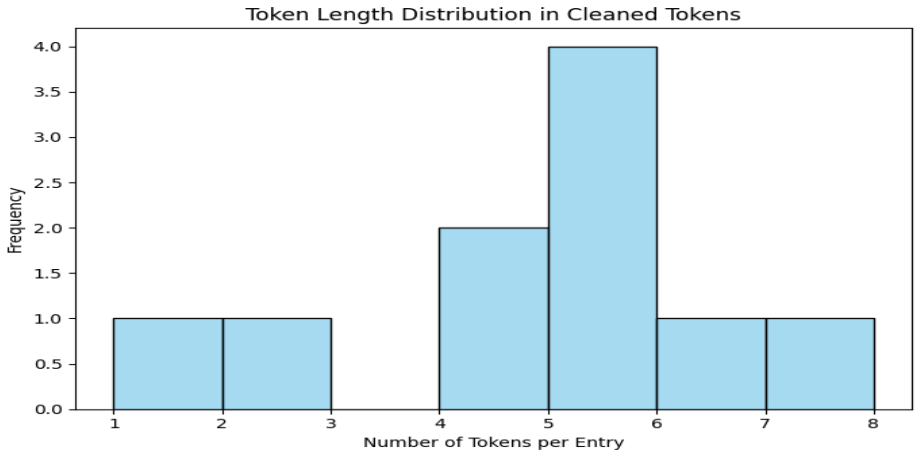


Fig. 3. Histogram of the Count of Tokens in Cleaned Igbo Lexical Entries

**TABLE II**
**PROMPT-TO-OUTPUT EXAMPLES USING ”UCHICHI” IN IGBOT5**

| Task | Example Output |
|---|---|
| Definition | Uchịchị bụ oge dị mkpirikpi,ma bụ mmalite nke ọchịchị tupu anyanwụ pụta |
| Paraphrasing | Ụchịchị bụ oge na-adị ntakịrị, nke na-egosi mmalite ọchịchị tupu anyanwụ apụta |
| Translation | Uchịchị is the part of the day before noon or Daybreak or Early hours before sunrise |
| Sentence Completion | . . . bu mgbe oyi dị elu, tupu mmadụ amalite ọrụ ụtụtụ. |
| Sentence Usage | N’uchichi, ana-aga ụlọ akwụkwọ |

### *C. Comparative Analysis with Existing Approaches*

IgboT5 represents a significant advancement over previous Igbo language models, particularly when compared to IgboBERT’s 84M parameters and limited NER-focused applications. Unlike IgboBERT’s encoder-only architecture, IgboT5’s text-to-text framework enables generative tasks that are crucial for language preservation and education. The model’s ability to handle diverse tasks simultaneously contrasts favorably with task-specific models that require separate training for each application.

The multilingual T5 foundation provides IgboT5 with crosslingual transfer capabilities that were absent in previous monolingual approaches. This is particularly valuable for Igbo, where parallel data is scarce. The model can leverage knowledge from high-resource languages while maintaining Igbo-specific linguistic patterns, addressing the fundamental challenge of low-resource language processing.

### *D. Cultural and Linguistic Preservation*

One of the most significant contributions of IgboT5 lies in its potential for cultural preservation. The model’s ability to generate contextually appropriate usage examples demonstrates an understanding of cultural nuances that extend beyond mere lexical translation. For instance, the sentence completion tasks often incorporate culturally relevant scenarios that reflect traditional Igbo society and contemporary usage patterns.

The inclusion of technical terms like ”AI” (Ak Iheakambere) in the dataset demonstrates the model’s capability to handle modern vocabulary while maintaining linguistic authenticity. This dual approach of preserving traditional expressions while accommodating contemporary terminology is crucial for the evolution and vitality of the Igbo language in digital spaces.

**TABLE III**
**PROMPT-TO-OUTPUT EXAMPLES USING ”UCHICHI” AND ”ABD” IN IGBOT5**

| Word | Task | Example Output |
|---|---|---|
| Uchichi | Definition | Uchichi b oge tt ma b mmalite bch tupu anyanw pta. |
| | Paraphrasing | Mbido tt, oge chi gbagburu, ma b mgbe chi chptara wa. |
| | Translation | Uchichi is the part of the day before noon. |
| | Completion | ... b mgbe oyi d elu, na mmad na-eme ntrnd tupu r. |
| | Usage | N’uchichi, maka na-aga l akwkw mgbe chi ptara. |
| Abd | Definition | Abd b mkpr edemede ma b njik nke mkprokwu d na alfabet Igbo. |
| | Paraphrasing | Ihe mejuptara abd Igbo b mkprokwu nd d ka a, b, ch, d, wdg. |
| | Translation | The Igbo alphabet |

| | | |
|---|---|---|
| | Completion<br><br><br><br>Usage | has various letters<br><br>.... na-enyere aka mta, g na ide ass a nke ma.<br><br>maka mtara abd Igbo tupu ha amalite ide akwkw. |

### *E. Limitations and Challenges*

Despite its promising results, IgboT5 faces several limitations that warrant discussion. The dataset size of 4,052 entries, while substantial for initial development, remains limited compared to high-resource language corpora. This constraint may affect the model's performance on domain-specific tasks or rare linguistic phenomena. The manual annotation process, while ensuring quality, limits scalability and may introduce unconscious biases from the annotation team. Additionally, the current focus on standard Igbo may not adequately represent dialectal variations across different Igbo-speaking regions, potentially limiting the model's applicability to diverse Igbo communities.

The evaluation methodology, primarily based on qualitative examples rather than quantitative metrics, presents another limitation. Future work should incorporate standardized evaluation protocols using metrics such as BLEU scores for translation tasks and semantic similarity measures for definition generation.

### *F. Implications for Low-Resource Language Processing*

IgboT5's development methodology offers valuable insights for other low-resource African languages. The success of adapting multilingual T5 for Igbo demonstrates the viability of transfer learning approaches for indigenous languages with limited digital resources. The careful balance between linguistic authenticity and practical applicability provides a template for similar endeavors in other African languages.

The text-to-text framework's flexibility proves particularly advantageous for languages with complex morphological structures like Igbo. The unified approach eliminates the need for task-specific architectures, reducing development overhead while maintaining performance across diverse applications. This efficiency is crucial for resource-constrained environments typical of low-resource language projects.

### *G. Educational and Societal Impact*

The potential applications of IgboT5 extend beyond academic research into practical educational tools. The model's definition generation capabilities could support digital dictionary development, while its paraphrasing abilities could assist in creating educational materials at various complexity levels. The translation functionality, though requiring further refinement, could facilitate cross-linguistic communication and content localization. For diaspora communities, IgboT5 represents a technological bridge to their linguistic heritage. The model's ability to generate culturally contextual content could support language learning initiatives and cultural preservation efforts in diaspora settings where traditional transmission methods may be compromised.

## 5. FUTURE WORK DIRECTIONS

1) The dataset will be expanded to other domains such as health, education, etc.
2) Integration of dialect variations
3) Building of bilingual and multilingual Igbo models
4) Deployment of the model in real-time as digital learning tools
5) Evaluation of IgboT5 on tasks such as summarization, question answering, and named entity recognition

## 6. CONCLUSION

This work presents IgboT5, an enhancement of African low-resource corpora specifically for the Igbo language using the Text-to-Text Transfer Transformer approach. Through the creation of a high-quality Igbo dataset containing 4,052 entries with proper linguistic annotations, we have laid the foundation for advanced NLP applications in the Igbo language. The preprocessing pipeline including normalization, tokenization, and task framing demonstrates the potential for applying state-of-the-art transformer models to low-resource African languages. The results show promising applications in definition generation,

paraphrasing, translation, and context completion tasks. This work contributes significantly to the advancement of African language NLP and opens new opportunities for future research in indigenous language processing.

## REFERENCES


[1] Odirichukwu, J. C and Nnamdi, R. N. Web-based igbo thesaurus with real-time retrieval, Journal of Computer Science Engineering and Software Testing, 1–8, 2022.

[2] Sawicki, J., Ganzha, M. and Paprzycki, M., 2023. The state of the art of natural language processing—a systematic automated review of NLP literature using NLP techniques. Data Intelligence, 5(3), 707-749.

[3] Mussandi, J. and Wichert, A., 2024, March. Nlp tools for african languages: Overview. In Proceedings of the 16th International Conference on Computational Processing of Portuguese 2, 73-82

[4] Chukwuneke, C., Ezeani, I., Rayson, P. and El-Haj, M., 2022, June. IgboBERT models: Building and training transformer models for the Igbo language. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 5114-5122).

[5] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A. and Raffel, C., 2021, June. mT5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies (483-498).

[6] Chi, Z., Dong, L., Ma, S., Huang, S., Singhal, S., Mao, X.L., Huang, H.Y., Song, X. and Wei, F., 2021, November. mT6: Multilingual pretrained text-to-text transformer with translation pairs. In Proceedings of the 2021 conference on empirical methods in natural language processing (pp. 1671-1683).

[7] Nagoudi, E.M.B., Elmadany, A. and Abdul-Mageed, M., 2021. AraT5: Text-to-text transformers for Arabic language generation. arXiv preprint arXiv:2109.12068.

[8] Chen, W.R., Abdul-Mageed, M. and Cavusoglu, H., 2021, June. IndT5: a text-to-text transformer for 10 indigenous languages. In Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas (pp. 265-271).

[9] Ekle, O.A. and Das, B., 2025. Low-Resource Neural Machine Translation Using Recurrent Neural Networks and Transfer Learning: A Case Study on English-to-Igbo. arXiv preprint arXiv:2504.17252.

[10] Ogundepo, O.J., Oladipo, A., Adeyemi, M., Ogueji, K. and Lin, J., 2022, July. AfriTeVA: Extending? small data? pretraining approaches to sequence-to-sequence models. In Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing (pp. 126-135).

[11] Odirichukwu, J.C., Chika-Ugada, P.K., Nnamdi, R.N., Odirichukwu, S.P.C., Ndigwe, C., Atolagbe, O.W., Dimoji, C., Njoku, O.A., Nwoke, J.C., Ekuma, G.O. and Durotola, I.T., 2025. Igbo Text Named Identity Recognition (NER) System using Natural Language Processing Algorithms: A Review. University of Ibadan Journal of Science and Logics in ICT Research, 15(1).