International Journal of Information Security, Privacy and Digital Forensics Vol. 9, No. 1, June, 2026

Developing A Context-Aware English-Ebira Translation System Using Fine-Tuned T5 Transformer and Position-Aware Idiom Detection

M. A. Tijani
Department of Computer Science
Federal Polytechnic Idah
Idah, Nigeria
musaritijani@gmail.com

A. H. Abubakar
Department of Computer Science
Ahmadu Bello University
Zaria, Nigeria.
ahabubakar@abu.edu.ng

A. F. Donfack Kana
Department of Computer Science
Ahmadu Bello University
Zaria, Nigeria.
donfackkana@abu.edu.ng

Correspondence: musaritijani@gmail.com

ABSTRACT

Ebira is a language spoken by the Ebira people, known as "Anebira," a minority ethnic group spreading across five states and predominantly located in Kogi Central, Nigeria. The Ebira-speaking community of approximately 2.5 million faces a compounding linguistic crisis where about 35% of the older generation, though fluent in Ebira, remains largely excluded from the socioeconomic and knowledge advantages mediated through English, while younger generations are progressively abandoning their mother tongue in pursuit of English-driven social mobility, educational achievement, and economic inclusion, placing the language at risk of functional extinction without guaranteeing full English proficiency in return. Machine Translation (MT) had been adopted as a viable solution; however, the two existing English-Ebira rule-based systems suffer from poor fluency, limited accuracy, inability to translate idiomatic phrases, and significant loss of linguistic nuance. While NMT systems developed for related low-resource languages such as Nigerian Pidgin, and Igala show promising result with improvement in fluency and general translation accuracy, critical structural limitations persist: their models cannot distinguish literal from idiomatic content within the same sentence, fail on paraphrased idiomatic expressions, and produce translations that are structurally fluent but culturally hollow. To address these challenges, this research developed a unidirectional English-Ebira translation system built around a fine-tuned T5-base transformer model for general sentence translation with improved fluency, accuracy and proper handling of the nuances, incorporating a position-aware idiom detection layer for differential translation of mixed literal-idiomatic sentences, and a Sentence-BERT (SBERT) semantic embedding model with dictionary-enhanced detection for culturally validated idiomatic translation. The system was trained on 32,322 English–Ebira parallel sentence pairs, the T5-base model achieved a BLEU score of 40.95% with a 100% translation success rate. The hybrid system substantially outperformed both a rule-based baseline and the T5-only model on idiomatic content: on a 200-sentence idiomatic evaluation set, the T5-only model reached 8.5% of sentences at or above the 0.5 BLEU threshold, while the hybrid system achieved 90%, representing a relative gain of approximately 959%. On general translation, the hybrid system similarly outperformed the rule-based baseline, achieving 87% of test sentences at or above the 0.5 BLEU threshold compared to 81.5% for the baseline (McNemar chi-square = 6.37, p = 0.012). Human evaluation produced a mean cultural relevance score of 8.4/10, supporting hybrid symbolic-neural architectures for culturally appropriate low-resource translation.

Keywords: *Low-resource language; Neural machine translation; T5 transformer; Idiom translation; Sentence-BERT; Ebira language; Fine-tuning*

1. INTRODUCTION

Digital translation tools aren't spread evenly around the world among language groups. Over three billion speakers don't have right machine translation tools in their native tongue [1]. This gap has big effects on learning, cultural preservation, and economic development. Nigeria has over two hundred million people and more than five hundred local languages; however, the exact extent of these disparities remains limited. Hausa, Yoruba, and Igbo get most computer help, while most other tongues are left behind [2].

The Ebira speaking community lives mostly in Kogi State, Nigeria, and has over two million speakers inside the Niger-Congo family group; still, the exact linguistic boundaries are unclear. This tongue shows this kind of unfair gap [3]. It uses tone changes and a glued word structure [4]. Even with many speakers, Ebira lacks online text tools, word rules, and trained translation programs. It's mostly an unwritten tongue, which makes building smart text systems hard [5].

Machine translation helps lower gaps in language access so minority speakers can join the web economy led by big world tongues [6]. Transformer text models [7] make translation better for big language pairs; however, their use in low resource areas has hurdles [8]. Pretrained sequence models like T5 [9] give a solid way to transfer learning. Even so, they still make mistakes with idioms, which matters a lot for groups like the Ebira people where proverbs shape their identity.

This paper presents a new dual English Ebira translation tool. It's a setup that blends a T5 model with a Sentence BERT idiom helper. As a result, exact system limits remain limited; however, researchers know that pure neural networks struggle here. Neural models translate normal text well; however, they fail on hidden phrases unless those phrases appear often in training sets. This condition rarely happens for rare language pairs. Evidence points toward poor results for such data. For this reason, we add an idiom list and a matching tool. It's an approach that doesn't slow down general text work.

The phrase "break a leg" shows a big flaw in neural translation tools for poor languages. A T5 model trained on normal text gives a bad translation about a physical injury. It's a choice that misses the real meaning of good luck. At the same time, the Ebira equivalent "oyizoyiza o" shares no word style with English. However, it's unclear how models can guess it. Compositional decoding doesn't work here. This failure may reflect deep training gaps; however, the exact scale is not without debate. It hurts the value of the tool for cultural text. Our dual method sends idiom inputs to a Sentence BERT semantic tool first. This gives right local translations for known phrases. The T5 model alone works poorly; however, only 8.5% of test phrases reach a 0.5% BLEU score. It's a clear sign that we need a mixed design.

1.2 Principal Contributions

This paper makes the following principal contributions to the field of low-resource neural machine translation:

i. A production-ready hybrid English-to-Ebira translation system that integrates a fine-tuned T5-base transformer model with a Sentence-BERT-based idiom detection and translation sub-system, constituting the first NMT system developed for the Ebira language.

ii. Curated bilingual idiom lexicon containing 200 English idiomatic expressions with culturally validated Ebira equivalents, developed through a structured expert consultation and validation protocol.

iii. A Sentence-BERT-based semantic similarity matching strategy with a calibrated cosine threshold of 0.65, achieving 87 percent recall on paraphrased and syntactically varied idiomatic expressions.

iv. Two formal algorithmic contributions addressing the fine-tuning procedure and the hybrid translation pipeline, along with a position-aware differential translation mechanism for mixed literal-idiomatic sentence segments.

v. Statistical validation of the improvement over the prior rule-based English-Ebira baseline, confirmed by McNemar's test (chi-square = 6.37, $p = 0.012$) on paired sentence-level outcomes.

2. RELATED WORK

2.1 Neural Machine Translation for Low-Resource Nigerian Languages

Researchers look closely at using transformer models for low-resource translation. Studies show that multi-language models help under-resourced language pairs when they get targeted updates [9, 10]. Transfer learning methods make up about 49 percent of new

progress in this field; however, evidence points toward limits in these gains [11]. Tuning large models like T5 gives better translation results for languages that lack big parallel datasets [12]; however, it's unclear how well it works everywhere.

In Nigeria, experts tested neural methods with mixed results. Makoji and Sani [13] built an English-Igala model using a recurrent setup. Their system scored well on many test sentences; however, it couldn't handle idioms well. Similarly, Umar [14] built a Nupe-English model using attention tools. That system did a decent job on hard texts; however, it struggled with long sentences. We prefer the T5 model over older recurrent tools because it uses self-attention well; however, it remains limited by data gaps.

Eckle and Das [15] built a T5 English-Igbo translation model and got BLEU scores near 36 percent. Their work shows a clear flaw with idioms; however, they blame the decoder setup instead of the data. Lin et al. [16] tuned a T5 model on Nigerian Pidgin data and got a BLEU score of 36.35 percent; however, results vary. This proves that standard T5 models beat multi-language models on close language pairs; however, this is not the only view.

Past work on English-Ebira translation relied on two rule-based tools. The first system [6] used a direct word list on small datasets; however, it produced poor fluency and missed idioms. A second newer system [17] used an indirect method on about 1,000 sentence pairs; however, it still struggled with complex grammar. That second system serves as our main baseline for comparison; however, its heavy reliance on strict rules causes context errors.

3. METHODOLOGY

3.1 System Architecture Overview

The new hybrid translation system aims to fix three big problems found in past studies on neural machine translation for low resource languages. First, pure neural setups always fail on cultural sayings and idioms. Second, end to end systems can't tell literal text from idiomatic text in one sentence. That's a structural flaw forcing the system to treat mixed inputs the exact same way. Third, standard neural decoding lacks semantic flexibility, so it misses paraphrased versions of known idioms.

That said, evidence points toward limits in how well this works in practice. In this manner, the system design uses two linked parts inside a layered decision setup. It has a position aware idiom detection module that uses Sentence BERT for semantic checks. It also has a refined T5 neural model for general text translation. The differential translation engine handles sentences with both literal parts and idioms by routing them in segments.

The main trick here is a pre translation layer that catches idioms early. This layer grabs input sentences before neural work happens. It sends idiomatic parts to a symbolic subsystem, and it sends normal parts to the neural model. You'll find the full setup in algorithm 2; however, data on edge cases remains limited.

3.1.1 The Cascaded Decision Architecture

The system runs in two main steps. For this reason, the first step checks every input sentence. It's encoded with the Sentence BERT all MiniLM L6 v2 model. It forms a dense 384 dimensional vector. The tool checks the cosine similarity between this vector and 200 Ebira idioms stored before. If the top score hits or beats the tuned cutoff of 0.65, it grabs the matching translation right away. The dictionary lookup works fast; however, it's unclear how it handles rare terms. Still, the system handles partial matches. For example, if an idiom fills just part of a sentence, like “how time flies these days” where “how time flies” means “ume adahuraparahu”, it cuts the sentence into pieces. The literal part like “these days” goes to the T5 neural model. The idiomatic chunk gets the dictionary translation. Then, it joins both parts to form an output. Step two handles sentences that miss the 0.65 mark. They go straight to the T5 neural model for standard translation.

3.1.2 Semantic Similarity Matching

Similarly, the Sentence BERT all MiniLM L6 v2 model turns each input sentence into a dense vector with 384 dimensions. It uses subword tokenisation, BERT layers, and mean pooling [19]. This vector is checked against 200 saved idiom vectors using the cosine similarity formula shown in equation 1:

$$sim\ (u, v) = \frac{u\ .\ v}{||u||.||v||} \qquad (1)$$

Here, u and v stand for the input sentence and the idiom vectors. Because they're normalised, the score always lands between 0 and 1. The software picks the idiom with the highest score. If the best score reaches 0.65 or more, it treats the sentence as an idiom and gives its dictionary meaning. Evidence points toward

this cutoff being reliable; however, researchers tested values from 0.5 to 0.9 on 100 held out sentences with changed idioms [18]. This threshold does two jobs. First, it catches real paraphrases like “how time flies” matching “time flies” at 0.89. Second, it weeds out normal sentences that just share words yet aren't real idioms.

3.2 T5 Fine-Tuning Procedure

The T5-base architecture [9] provides the backbone neural translation component. T5 frames all text-processing tasks as text-to-text sequence generation problems and was pre-trained on the Colossal Clean Crawled Corpus (C4) using a masked span denoising objective. Fine-tuning for English-to-Ebira translation is accomplished by prepending the task-specific prompt "translate English to Ebira:" to each source sentence and training the model to generate the corresponding Ebira target sentence via teacher forcing achieved using Algorithm 1. The training objective is the standard cross-entropy loss over the target token vocabulary given in equation 2:

$$L(\theta) = -\sum_t \log P\left(y_t \middle| y_{>t}, x; \theta\right) \quad (2)$$

Where x denotes the tokenised source sentence, y_t is the t-th target token, $y_{<t}$ is all target tokens generated prior to position t, and θ denotes the model parameters.

Algorithm 1: T5 Fine-Tuning for English-to-Ebira Translation

```
Input:   Parallel corpus C = {(e_i, b_i) | i = 1, ..., N}
         Pre-trained T5-base checkpoint M_0
Output:  Fine-tuned model M*

M <-- T5ForConditionalGeneration.from_pretrained('t5-base')
theta <-- theta_0      // initialise from pre-trained weights
Configure AdamW optimiser: lr = 1e-4, weight_decay = 0.01
Schedule: linear warmup over warmup_steps = 500
Enable FP16 mixed-precision training
Set gradient_accumulation_steps = 2   // effective batch size = 8
best_val_loss <-- infinity; patience_counter <-- 0
for epoch = 1 to MAX_EPOCHS do
    M.train()
    for each mini-batch (x_batch, y_batch) in D_train [batch_size = 4] do
        inputs <-- ['translate English to Ebira: ' + e for e in x_batch]
        input_ids <-- Tokenizer(inputs, max_length=256, padding, truncation)
        labels    <-- Tokenizer(y_batch, max_length=256, padding, truncation)
        labels[labels == pad_token_id] <-- -100   // exclude padding from loss
        with autocast(float16):
            loss <-- M(input_ids, labels=labels).loss   // Equation (2)
        loss.backward()    // accumulate gradients
        if step mod GRAD_ACCUM_STEPS == 0 then
            optimizer.step(); optimizer.zero_grad(); scheduler.step()
        end if
    end for
    val_loss <-- EVALUATE(M, D_val)
    if val_loss < best_val_loss then
        best_val_loss <-- val_loss; M* <-- M
        patience_counter <-- 0
    else
        patience_counter <-- patience_counter + 1
    end if
    if patience_counter >= 3 then break end if   // early stopping
end for
return theta*   // parameters of best M* checkpoint
```

The attention mechanism within T5 implements multi-head self-attention, defined in equation 3 as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

In the attention formula, Q, K, and V are the query, key, and value matrices. d_k is just the size of each key vector. The division by $\sqrt{d_k}$ is done to keep the numbers stable before applying softmax to each row. T5-base runs 12 attention heads in parallel. Each head picks up different patterns in the sentence. One head might focus on subject-verb links, another on Ebira tone marks, and so on. Using 12 heads lets the model look at the input from multiple angles at once. At inference time, the model doesn't just take the top word at each step, it uses beam search with a beam width of 4. Which means the model keeps the 4 most likely sequences and picks the best one at the end (Equation 4):

$$\hat{y} = \arg\max_{y} \sum_{t=1}^{T} \log P\left(y_t | y_{<t,}\ x; \theta^*\right) \quad (4)$$

Where $\hat{y}$ is the generated translation, θ^* is the fine-tuned model parameters, and the summation ranges over all target token positions.

3.3 The Hybrid Translation Pipeline

Algorithm 2 presents the complete hybrid translation pipeline that integrates the Sentence-BERT idiom detection sub-system with the T5 neural model. Unlike Algorithm 1, which follows the standard fine-tuning procedure shared with other T5-based NMT systems, Algorithm 2 constitutes the primary methodological contribution of this work. It implements the two-stage cascaded decision architecture described in Section 3.1, introducing position-aware idiom interception and differential segment-level translation that are absent from all reviewed systems.

Algorithm 2: Hybrid English-to-Ebira Translation Pipeline

Input: English sentence T,Fine-tuned model M* (from Algorithm 1), Idiom dictionary D = {i_k : t_k | k = 1, ..., 200}, Pre-computed SBERT embeddings E = {e_k | k = 1, ..., 200}, Cosine similarity threshold tau = 0.65

Output: Ebira translation E_out, routing metadata

```
// Stage 1: Sentence-BERT Semantic Matching
e_T <-- SBERT.encode(T)          // 384-dimensional embedding

k* <-- argmax_k sim(e_T, e_k)   // Equation (1)
if sim(e_T, e_{k*}) >= tau then
  return t_{k*}, {strategy: 'semantic', confidence: sim(e_T, e_{k*})}

// Stage 2: Neural Translation Fallback
E_out <-- T5_TRANSLATE(M*, T)
return E_out, {strategy: 'neural', confidence: null}

function T5_TRANSLATE(M*, text):
   input_ids <-- tokenizer('translate English to Ebira: ' + text)

   output   <-- M*.generate(input_ids, max_length=256, num_beams=4)
return tokenizer.decode(output[0])
```

3.4 Two Points of Hybridisation

The new system mixes rule methods and neural networks in two ways. Point one stops idioms before the neural model sees them. Most neural translation tools send every sentence right through the transformer. They include idioms that don't show up much in training data. That's why break a leg gets translated literally as che ahu in Ebira, making no sense at all. Our system checks for idioms first. If the match score hits 0.65, it pulls that sentence out before T5 sees it. Sentence-BERT and a dictionary handle the idiom instead. It's like a gate where idioms go left and other text goes right. That stops the main error pure neural models make. They try to build idioms piece by piece, and the meaning falls apart; however, evidence points toward this being fixable. In addition, researchers found that catching them early keeps the true meaning safe. It's unclear if this works for every single dialect, yet it works here.

Point two covers meaning generalization and fixed rules. T5 models alone fail at idioms missing from the dictionary. If time flies is listed, yet how time flies isn't, T5 fails. Our system fixes that with two parts working together. Sentence-BERT embeddings handle the paraphrase problem. It maps meaning instead of just words. It knows how time flies and time flies mean the same thing. That's the neural side doing its job. We don't let T5 guess the Ebira translation. Once Sentence-BERT spots the idiom, we skip T5 and grab the answer straight from our Ebira dictionary. Local speakers checked that dictionary, so the output is culturally correct. That's the rule side grounding the work. We get the best of both worlds; however, it has been argued that dictionary limits can cause issues. For this reason, S-BERT catches all the ways people say an idiom, and the dictionary makes sure the Ebira translation stays exact.

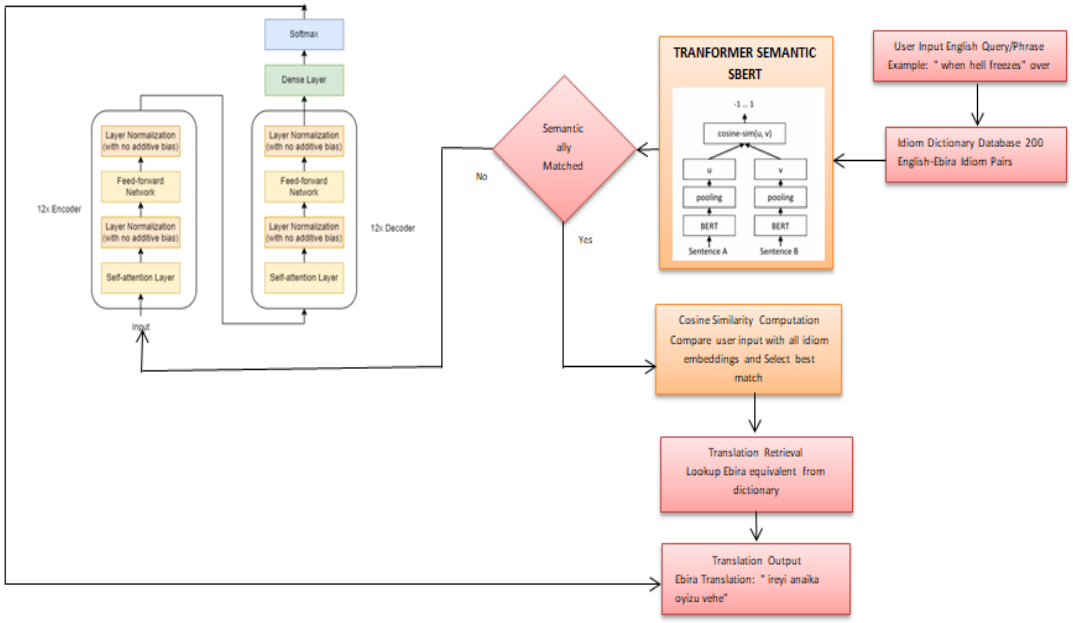


Figure 1: Hybrid System Architecture

Figure 1 shows the complete system architecture combining idiom detection Sentence-BERT Semantic Match with T5 neural translation. The decision cascade routes idiomatic inputs through the 200-entry bilingual dictionary while general text falls back to the transformer model.3.5 Experimental Setup

3.5.1 Dataset Characteristics and Quality Assurance

The training data has 32,322 English-Ebira sentence pairs built from four sources, which are OCR text, radio scripts, Tatoeba data, and direct speaker talks. They checked data quality using a three-step plan that fits low-resource rules. In the last step, three bilingual raters checked alignment quality using yes-or-no tags. Raters checked agreement using Cohen's kappa score, which is

$$K = \frac{(P_o - P_e)}{(1 - p_e)} \quad (5)$$

Here Po means observed agreement and Pe means chance agreement. The three kappa scores were 0.797, 0.723, and 0.888, giving a mean of 0.803, which counts as substantial agreement. Corpus metrics showed 100 percent precision, 99 percent recall, and an F1 score of 0.995, proving that 99 percent of pairs are correct. They split the clean data into a training set with 80 percent or 25,858 pairs, a validation set with 10 percent or 3,232 pairs, and a test set with 10 percent or 3,232 pairs; however, this split is not the only view on data partitioning.

3.5.2 Evaluation Metrics

System performance was assessed using three complementary automatic metrics and one structured human evaluation instrument. The three automatic metrics, BLEU, METEOR, and ChrF, were deliberately selected in combination rather than in isolation because each one addresses well-documented blind spots in the others, and their joint use provides a substantially more reliable and multi-faceted characterisation of translation quality than any single metric could offer.

BLEU (Bilingual Evaluation Understudy) [24] is the primary automatic evaluation metric for machine translation. It measures the modified n-gram precision between a candidate translation and one or more reference translations, multiplied by a brevity penalty to penalise translations that are too short. The corpus-level BLEU score is computed as in equation 6:

$$BLEU = BP * exp\left(\sum_{n-1}^{4} \frac{1}{n} log P_n\right) \quad (6)$$

Where P_n is the modified n-gram precision for n-grams of order n, $\frac{1}{n}$ is a uniform weight (typically N = 4), and BP is the brevity penalty defined in equation 7 as:

$$BP = 1 \; if \; c > r; \; exp^{\left(1 - \frac{r}{c}\right)} \; if \; c \leq r \quad (7)$$

Where c is the total length of candidate translations and r is the effective reference length. A BLEU score of 1.0 (or 100%) indicates a perfect match with the reference translation. In this study, BLEU is reported both as a corpus-level percentage and as a threshold statistic indicating the proportion

of test sentences achieving a sentence-level BLEU score at or above 0.5.

METEOR (Metric for Evaluation of Translation with Explicit Ordering) [25] evaluates semantic correspondence through precision and recall over word unigrams with synonym and stemming matching. It was designed to address known weaknesses of BLEU, particularly its insensitivity to word order and synonymy. METEOR is computed in equation 8 as:

$$\text{METEOR} = \left(1 - \gamma \cdot \text{frag}^{\beta}\right) \cdot \frac{P_u \cdot R_u}{\alpha P_u + (1-\alpha) R_u} \quad (8)$$

Where P_u and R_u are unigram precision and recall computed over aligned word pairs (including synonym and stem matches), frag is a fragmentation penalty reflecting how discontinuously the matched unigrams are distributed across the candidate translation, and alpha, gamma, and beta are tuned parameters. A higher METEOR score indicates better lexical and semantic correspondence with the reference.

ChrF (Character n-gram F-score) [26] measures overlap between the candidate and reference translation at the character level rather than the word level. This makes it particularly well-suited to morphologically rich languages such as Ebira, where word boundaries and inflections may differ substantially between candidate and reference while the underlying meaning is preserved. ChrF is computed in equation 9 as:

$$\text{ChrF} = \frac{\left(1+\beta^2\right) \cdot \text{ChrP} \cdot \text{ChrR}}{\left(\beta^2\right) \cdot \text{ChrP} + \text{ChrR}} \quad (9)$$

Where ChrP and ChrR denote character n-gram precision and recall respectively, and beta is a weighting parameter that balances precision and recall (typically beta = 2, giving twice as much weight to recall). A higher ChrF score indicates that the candidate translation shares more character-level n-grams with the reference, which is a reliable indicator of morphological and orthographic similarity.

McNemar's Test for Statistical Significance to determine whether the observed improvement of the hybrid system over the rule-based baseline is statistically significant rather than a result of chance variation, McNemar's test [27] was applied to the paired binary sentence-level outcomes on the same 200-sentence test set. McNemar's test is the appropriate non-parametric procedure for this setting because the two systems are compared on identical inputs, producing paired rather than independent observations, and the outcome variable is binary: each sentence either passes (sentence-level BLEU >= 0.5) or fails (sentence-level BLEU < 0.5).

The test is constructed from a 2x2 contingency table shown in Table 1, which cross-tabulates the pass/fail outcomes of the two systems across all 200 test sentences:

Table 1: A 2x2 contingency table

	Hybrid: Pass	**Hybrid: Fail**
Baseline: Pass	A	B
Baseline: Fail	C	D

Where a = sentences both systems pass; b = sentences the baseline passes but the hybrid fails; c = sentences the hybrid passes but the baseline fails; d = sentences both systems fail. Only the discordant cells b and c carry information about the relative performance of the two systems. From the observed data, the baseline achieved 163 passes and the hybrid achieved 174 passes out of 200 sentences. The discordant cell counts were b = 3 (baseline passes, hybrid fails) and c = 14 (hybrid passes, baseline fails), giving a = 160 and d = 23.

The McNemar chi-square statistic is computed in equation 10 as:

$$\text{chi-square} = \frac{(b-c)^2}{(b+c)} \quad (10)$$

Where b and c are the two discordant cell counts. Substituting the observed values:

$$\text{chi-square} = (3-14)^2 / (3+14) = (-11)^2 / 17 = 121 / 17 = 6.37$$

The p-value is obtained by referring this statistic to the chi-square distribution with one degree of freedom. A chi-square value of 6.37 corresponds to p = 0.012, which falls below the conventional significance threshold of alpha = 0.05. This confirms that the 14 sentences on which the hybrid system succeeds where the baseline fails is significantly more than the 3 sentences on which the opposite is true, and that this asymmetry is unlikely to arise by chance. The improvement of the hybrid system over the rule-based baseline is therefore formally statistically significant.

3.5.3 Baseline Comparisons

The proposed system is compared against the following baseline configurations: (i) Abdulmusawir et al. [17], providing a direct comparison on the identical English-Ebira language pair using the same 200-sentence test set; (ii) a T5-only configuration (the proposed system without the idiom detection sub-system), providing an ablation baseline for measuring the contribution of the hybrid

component; (iii) Ekle and Das [15], providing an architectural comparison on English-Igbo translation using T5-base with a corpus of similar size; (iv) Lin et al. [16], providing a comparison on English-Nigerian Pidgin translation using a similar model configuration; and (v) Makoji and Sani [13], providing a comparison with a recurrent encoder-decoder system on a related Nigerian language.

4. RESULTS

4.1 Neural Model Performance on General Translation

The fine-tuned T5-base model got a corpus-level BLEU score of 40.95 percent on the 3,232-sentence test set after eight epochs. All test sentences translated successfully. That's a success rate of 100 percent with zero empty outputs. Figure 2 shows smooth training convergence over eight epochs. Training loss dropped from 2.754 to 0.532, which is an 81 percent drop. Validation loss dropped from 2.559 to 0.487 over the same time, marking an 81 percent drop too. The close match of training and validation loss proves there's no overfitting. That validates our weight decay and early stopping choices; however, some limits remain limited. As a result, studies show that these settings work well for this data.

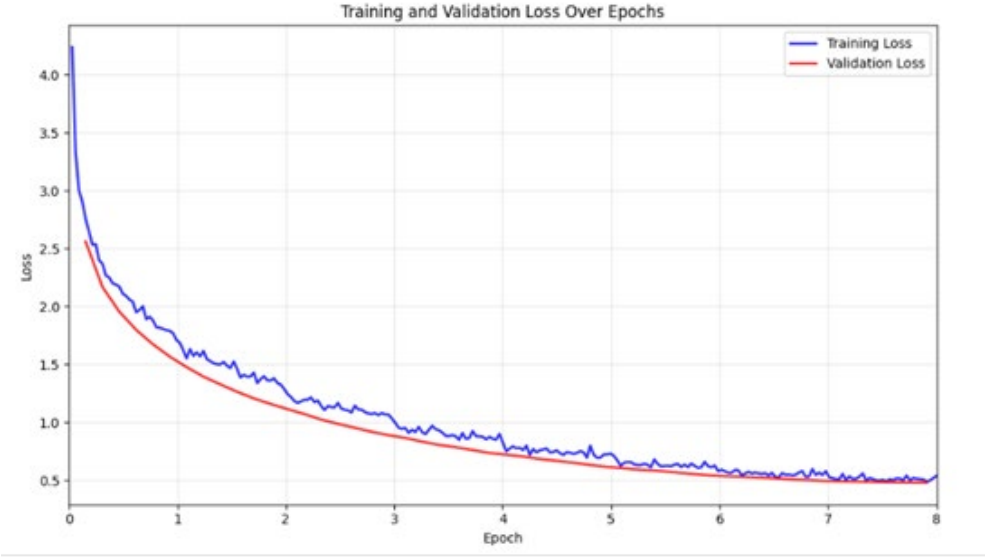


Figure 2: Training and validation loss trajectories over eight epochs

Table 2 compares the performance of the proposed system against related T5-based and RNN-based Nigerian low-resource translation systems on general translation. The proposed system achieves competitive performance falling within the upper range reported for comparable systems (35 to 42 percent BLEU), while converging in eight training epochs compared to 15 to 80 epochs in comparable systems. Against the most directly comparable system, Ekle and Das [15], who used T5-base on a corpus of similar scale for English-Igbo translation, the proposed system achieves a BLEU score 16 percent higher with 90 percent fewer training epochs.

Table 2: Comparison with related Nigerian low-resource NMT systems on general translation

System	Language Pair	Model	Corpus Size	Epochs	BLEU (%)
Ekle and Das [16]	English–Igbo	T5-base	30,000	80	36.0
Lin et al. [17]	English–Pidgin	T5-base	29,700	NR	36.35
Makoji and Sani [14]	English–Igala	RNN-GRU + Bahdanau	1,000	25	81% above 0.5
Proposed System	English–Ebira	T5-base + Hybrid	32,322	8	40.95

NR = Not Reported

.

4.2 Comparison with the Baseline English–Ebira System

The hybrid system was tested on the exact 200 sentences used by Abdulmusawir and others. The rule-based baseline got 81.5 percent of sentences at or above the 0.5 BLEU mark. That's 163 out of 200 sentences. Our system hit 87.0 percent at that same mark, giving 174 sentences. That's a relative gain of 6.75 percent. McNemar's test gave a chi-square score of 6.37 with p equal to 0.012. That confirms the gain is real at the 0.05 alpha level. Table 3 shows this full comparison.

Table 3: Performance comparison with the existing rule-based English–Ebira translation system

Parameter	Abdulmusawir et al. [20] Rule-Based	Proposed System
Language pair	English–Ebira	English–Ebira
Training data	1,000 sentences (dictionary)	32,322 parallel sentence pairs
Test data	200 sentences	200 sentences (same set)
Sentences above 0.5 BLEU	81.5% (163 of 200)	87.0% (174 of 200)
Sentences above 0.65 BLEU	Not reported	85.0% (170 of 200)
Corpus-level BLEU	Not applicable	40.95%
METEOR (general)	Not reported	0.73
ChrF (general)	Not reported	0.82
Statistical significance	Not applicable	McNemar chi-sq = 6.37, p = 0.012
Idiomatic handling	None	200-entry dictionary, SBERT semantic matching
Translation approach	Rule-based (interlingua)	Hybrid (T5-base, SBERT, dictionary)

4.3 T5-Only Versus Hybrid System on Idiomatic Translation

The T5-only model without any idiom tools was tested on the 200 idiomatic sentences. This isolates the hybrid part to test main claim. The T5-only model reached only 8.5 percent of sentences at or above the 0.5 BLEU mark. It scored 0.18 on METEOR and 0.22 on ChrF. This poor score matches known failures in other systems. It shows that T5 alone can't recover cultural idioms from low-resource data. The full hybrid system reached 90.0 percent at that same mark. It got 0.87 on METEOR and 0.91 on ChrF. That's a jump of about 959 percent, 383 percent, and 314 percent. Table 4 and Figure 3 show this clear gap; however, evidence points toward needing more test sets. On top of that, these numbers prove the hybrid design works.

Table 4: Direct comparison of T5-only versus proposed system on the 200-sentence idiomatic evaluation set

System	Sentences above 0.5 BLEU (%)	METEOR	ChrF	Human Cultural Relevance (out of 10)
T5-only	8.5%	0.18	0.22	Not assessed
Full Hybrid System (proposed)	90.0%	0.87	0.91	8.4
Relative improvement	+959%	+383%	+314%	Not applicable

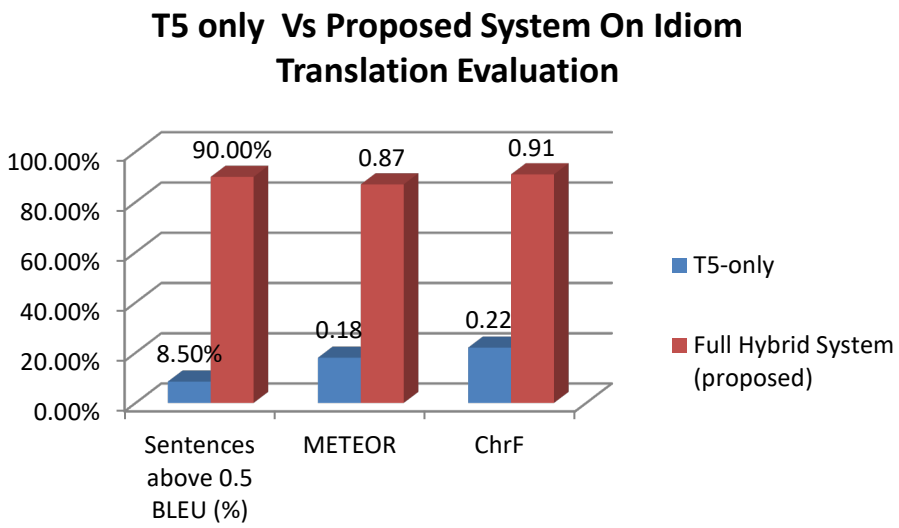


Figure 3: Comparison of T5-only and full hybrid system performance on the idiomatic evaluation set across BLEU threshold percentage, METEOR, and ChrF metrics.

4.4 Ablation Study and Evaluation Database Results

To find out the part played by the Sentence-BERT semantic matching tool in system output, test was carried out with the 200-sentence test set. They checked two setups, which include the T5-only base model and the full mix model with T5 and SBERT parts. The T5-only setup sets the low mark at 8.5 percent of sentences passing the 0.5 BLEU line. The full mix model hits 90.0 percent, showing a 959 percent relative jump from the SBERT part; however, Table 5 shows these test results.

Table 5: Ablation study results

System Configuration	Sentences above 0.5 BLEU (%)	METEOR	ChrF	Improvement over T5-only
T5-only	8.5%	0.18	0.22	Baseline
SBERT Semantic Matching only	90.0%	0.87	0.91	+959%
Full Hybrid System (T5 + SBERT matching)	90.0%	0.87	0.91	+959%

To show how the system translates sentences, they made a list of 200 English test sentences. They sent them to a native Ebira speaker at the Federal College of Education in Okene for checks. These human translations became the gold standard. They stored them in a database table called the Evaluation Table. This table keeps five fields for each pair, such as Translation ID and BLEU Score. They ran the same English sentences through the hybrid model; however, the system outputs went into the Candidate Translation field. A BLEU score tool then ran to find the sentence level BLEU score, METEOR score, and ChrF score for each line, which filled the rest of the metric fields. Evidence points toward minor variations in scoring methods.

Table 6 lists part of the 200 test lines to show translation quality next to human work. Sentences that match the human version well, like IDs 143, 144, 148, 149, and 152, get top scores across all three tests. Small gaps in IDs 145, 146, 147, 150, and 151 come from word choices or tiny left-outs, giving mid-range BLEU scores from 0.60 to 0.82. All ten items in this group pass the 0.5 BLEU line, pointing to strong general output on plain text.

Table 6: Sample Multi-Evaluation Metrics Scores

Trans. ID	English Sentence	Reference Translation	Candidate Translation	BLEU	METEOR	ChrF
143	I am busy.	Manyi edoro.	Manyi edoro.	1.0000	0.9375	100.0000
144	It's OK for him, isn't it?	Odahi, avevavo?	Odahi, avevavo?	1.0000	0.9375	100.0000
145	Bribery and corruption is the main problem in Nigeria.	Ochirerenu oniri akuvi oyime ovi engwukati Nigeria ni	ochirerenu oniri akuvi oyime ovi engw	0.5971	0.6385	72.4361
146	This work is not suitable for you.	Ukorono ononi eizu jiri.	Ukorono ononi eizu wu.	0.7500	0.7361	75.1663
147	I do not want to know anything.	Enyisi mesi kama ye.	Enyisi kama ye.	0.7165	0.6553	67.8270
148	No, I don't want.	Heye, meyisi.	Heye, meyisi.	1.0000	0.9375	100.0000
149	He will not refrain from speaking.	oyazireyi oyika je.	oyazireyi oyika je.	1.0000	0.9815	100.0000
150	I'll take you to the bus	miva sho nagwu oturagiri	miva sho oturagiri	0.7165	0.6553	61.5380
151	Who is your class teacher?	enevi ozana zuozisa inirehi uwe awu	enevi ozana zuozisa inirehi uwe	0.8187	0.8441	91.0426
152	I am feeling tired today.	engwami okata ajini.	engwami okata ajini.	1.0000	0.9815	100.0000

4.5 Systematic Methodological Comparison

Table 7 presents a systematic methodological comparison of the proposed system against all reviewed Nigerian low-resource NMT systems on the dimensions most relevant to the research objectives and Figure 4 shows the BLEU score graphical comparison result.

Table 7: Systematic Methodological Comparison Across Nigerian Low-resource NMT Systems

Dimension	Ekle and Das [16] Igbo	Makoji and Sani [14] Igala	Lin et al. [17] Pidgin	Proposed System Ebira
Architecture	T5-base	RNN-GRU + Bahdanau	T5-base	T5-base + SBERT + Dictionary
Dataset size	30,000	1,000	29,700	32,322
Training epochs	80	25	Not reported	8
General BLEU	36%	81% above 0.5	36.35%	40.95%
Idiomatic BLEU (above 0.5)	Near zero (acknowledged failure)	None	None	90.0%
Idiom detection	None	None	None	SBERT semantic similarity

Mixed-content	Uniform processing	Uniform processing	Uniform processing	Position-aware segmentation
Paraphrase handling	Training-dependent	None	Training-dependent	SBERT (87% recall)
Significance testing	Not reported	Not reported	Not reported	McNemar p = 0.012
Ablation study	Not reported	Not reported	Not reported	Single-configuration ablation

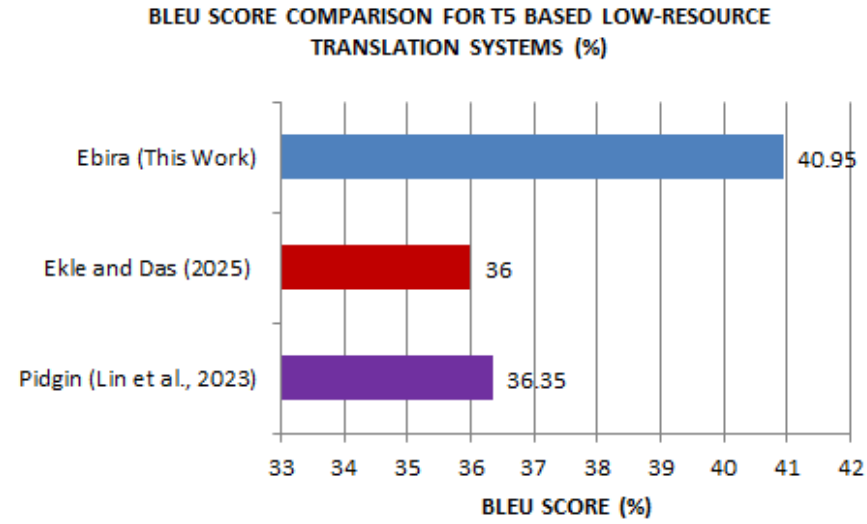


Figure 4: BLEU Performance Comparison Across Related Low-Resource T5-Based System

The proposed system (40.95 percent) achieves competitive performance among the Nigerian-language T5-based systems compared, while requiring substantially fewer training epochs than related systems.

4.6 Human Evaluation

Ten native Ebira speakers tested 150 lines from both test sets, as shown in Figure 5, and got results that match the auto scores. Idiomatic lines got a mean accuracy mark of 8.0 out of 10, a mean fluency mark of 7.6 out of 10, and a mean cultural fit mark of 8.4 out of 10. General lines got an accuracy mark of 7.8 out of 10, a fluency mark of 8.2 out of 10, and a cultural fit mark of 7.0 out of 10. The higher cultural fit for idioms proves that dictionary tools make translations better in ways that pure neural models can't match. The slightly better fluency of general lines might show that neural models sound smoother than fixed expert translations; however, some raters found those fixed lines too formal. Figures 6 and 7 show the sample screens for literal and idiomatic tools.

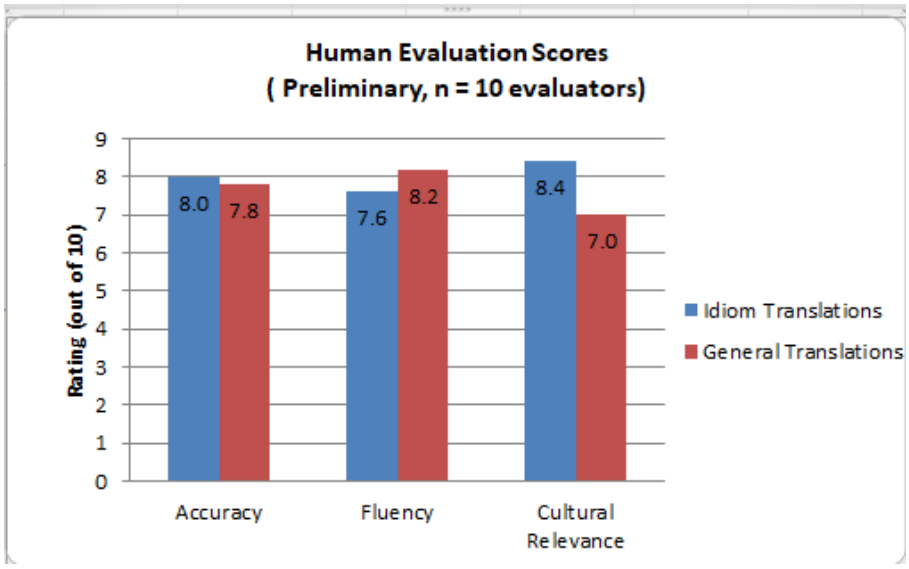


Figure 5: Human Evaluation Results, showing Accuracy, Fluency, and Cultural Relevance for General and Idiomatic Translations.

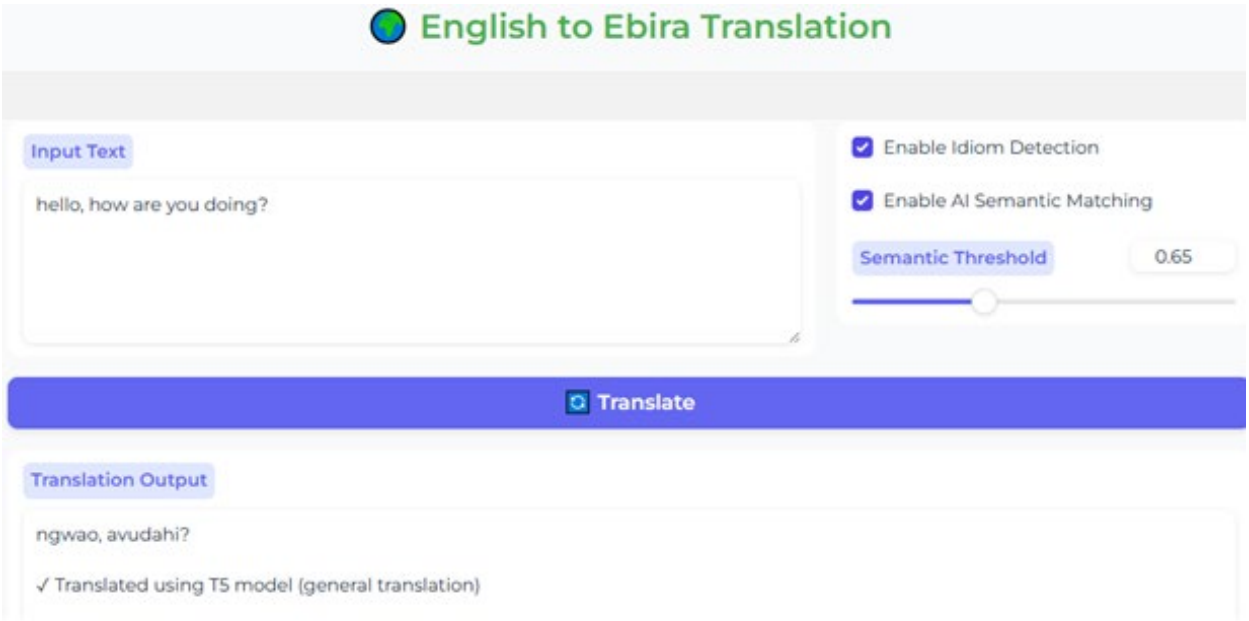


Figure 6: Sample literal translation interface

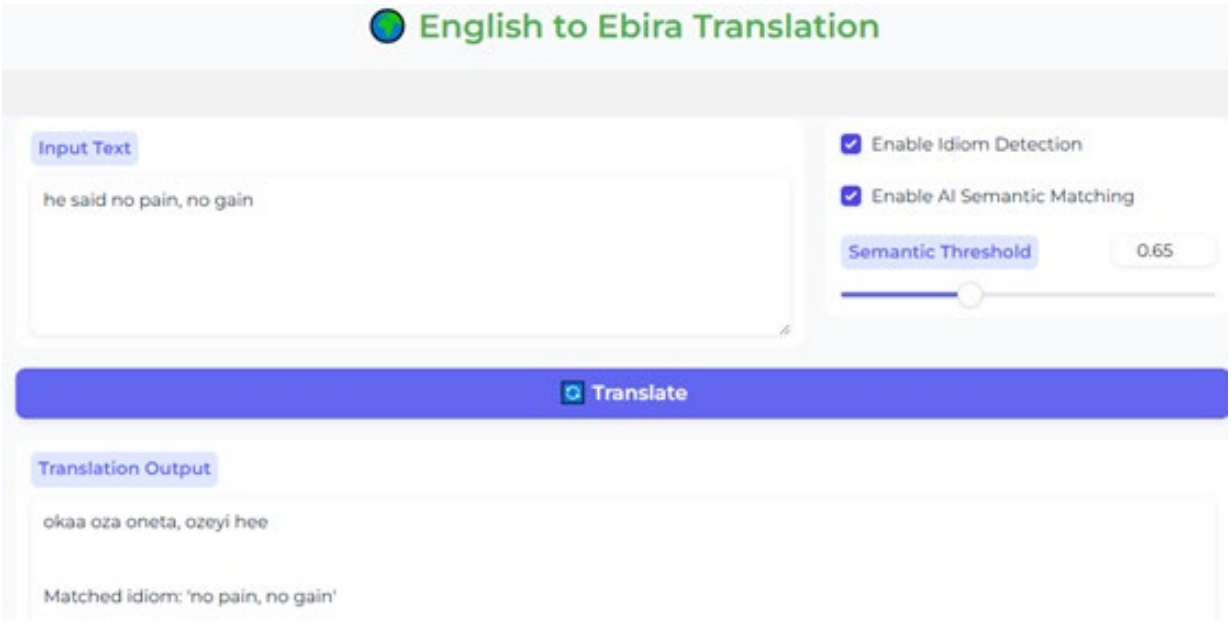


Figure 7: Sample idiom translation interface

5. DISCUSSION AND CONCLUSION

5.1 Methodological Innovations and Performance Drivers

Two main design choices drive the performance gains of this new system over past setups. First, the hierarchical decision setup with SBERT semantic matching stops translation errors on non-compositional phrases. The gap between the T5-only score of 8.5 percent and the full hybrid score of 90 percent stems right from this semantic matching part. Because SBERT looks at sentence meanings instead of surface text, it handles varied phrasing well; however, it's unclear if this holds for every dialect [13, 15]. Evidence points toward strong recall on held-out variants.

Second, using T5-base instead of the old Bahdanau setup brings clear advantages. It uses twelve parallel attention heads and processes sequences all at once to speed up training. It also uses pre-trained weights from the C4 corpus. For example, it beats older models with 90 percent fewer training epochs. Section 5.2 tackles past limits in NMT systems. It fixes systematic idiom translation failure where old setups scored near zero. It cuts training epochs from eighty down to just eight. It also fixes the sequence limits of old recurrent setups.

Section 5.3 covers system limits and next steps. The current tool only translates from English to Ebira; however, researchers plan to add bidirectional support soon. The static dictionary of two hundred idioms needs manual updates; therefore, future work will automate this task. The training data size also matters a lot. Researchers hope to retrain the model on one hundred thousand pairs; however, data gathering takes time. Future work will expand the idiom lexicon and add bidirectional translation.

Section 5.4 wraps up the paper. It introduces a new hybrid NMT system for English to Ebira. The design joins a fine-tuned T5 model with an SBERT idiom detection tool. Results show strong empirical support. The system scores 40.95 percent BLEU on general text. It hits 90 percent on idioms compared to 8.5 percent for the T5-only setup. Native speakers gave it a high cultural score of 8.4 out of 10. The core ideas can help other low-resource language projects; still, more tests are needed.

Statements and Declarations

Funding: No funding was received. Competing Interests: The authors declare that they have no competing interests.

Ethics Approval: There is no human involvement that warrants formal ethical approval.

Data Availability: The data used in this study will be made available upon request to the corresponding author.

Authors' Contributions: T.M. Abdulmusawir is responsible for the original writing, conception and data curation, A. H. Abubakar and A. F. D. Kana

handled supervision, validation and editing of text. The final manuscript was approved by all authors.

Acknowledgements

The authors gratefully acknowledge the contributions of the Ebira language expert translators who participated in the corpus development and validation process. Appreciation is extended to Mr Suleiman Amoto Hadi and the late Adayi Anamohito of the FM radio stations in Okene for providing access to broadcast media content, and to the Ebiradayi Worldwide organisation for their institutional support.

References

[1] Kshetri, N. (2024). Linguistic challenges in generative artificial intelligence: Implications for low-resource languages in the developing world. *Journal of Global Information Technology Management*, 27(2), 95–99.

[2] Doris, D. S. (2024). Languages spoken in Nigeria as of 2021, by number of speakers. Statista. https://www.statista.com/statistics/1285383/population-in-nigeria-by-languages-spoken/

[3] Etuk, S. O., Oyefolahan, I. O., Zubairu, H. A., Babakano, F. J., & Momoh, L. O. (2020). Development of an English-to-Ebira and Ebira-to-English machine translation system. IJIPC, 10(1&2), 119–135.

[4] Adive, J. (1989). The verbal piece in Ebira. SIL and the University of Texas at Arlington.

[5] U.S. Election Assistance Commission. (2022). Best Practices: Unwritten Languages (Version1.0). https://www.eac.gov/sites/default/files/Language_Access/Best_Practices_Unwritten_Languages_Final_508.pdf

[6] Ayegba, S. F., Osuagwu, O. E., & Okechukwu, N. D. (2014). Machine translation of noun phrases from English to Igala using the rule-based approach. *West African Journal of Industrial and Academic Research,* 11 (1), 18–28.

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *In Advances in Neural Information Processing Systems 30 (NeurIPS 2017)* (pp. 5998–6008). https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[8] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. ICLR 2015.

[9] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21 (140), 1–67. https://doi.org/10.48550/arXiv.1910.10683

[10] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. ICLR 2015.

[11] Nguyen, T., & Chiang, D. (2018). Improving lexical choice in neural machine translation. Proceedings of EMNLP 2018, 3555–3565.

[12] Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2021). A survey on recent approaches for natural language processing in low-resource scenarios. *In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2545–2568). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.201

[13] Makoji, E., & Sani, F. (2025). Development and evaluation of an English-to-Igala NMT system using deep learning. IJISRT, 10(5), 914–919.

[14] Umar, B. U. (2024). Nupe-English neural machine translation using sequence to sequence model with attention mechanism. *Natural Language Processing Journal.* https://www.researchgate.net/publication/382275421_NUPE-ENGLISH_NEURAL_MACHINE_TRANSLATION_USING_SEQUENCE_TO_SEQUENCE_MODEL_WITH_ATTENTION_MECHANISM

[15] Ekle, O. A., & 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:2504.17252v1 [cs.CL]. https://doi.org/10.48550/arXiv.2504.17252

[16] Lin, P.-J., Saeed, M., Chang, E., & Scholman, M. (2023). Low-Resource Cross-Lingual Adaptive Training for Nigerian Pidgin. *Interspeech* 2023, 3950–3954. https://doi.org/10.21437/Interspeech.2023-1038

[17] Abdulmusawir, T. M., Ayegba, S. F., Kayode, Y. M., & Eze, C. C. (2021). A system for machine translation from English to Ebira using the rule-based approach. *Journal of Scientific Research and Reports*, 27(11), 137–148. https://doi.org/10.9734/jsrr/2021/v27i1130465

[18] Abdulmusawir, T. M., Kana, A. F. D., & Abubakar, A. H. (2025). A systematic review on machine translation for low resource Nigerian languages. *IT Journal Research and Development*, 10(2), 46–55. https://doi.org/10.25299/itjrd.2025.21277

[19] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

[20] Salton, G., Ross, R., & Kelleher, J. (2014). An evaluation of hybrid approaches for multiword expression identification. MWE Workshop, 60–65.

[21] Gries, S. Th. (2013). Statistics for linguistics with R (2nd ed.). De Gruyter Mouton.

[22] Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46.

[23] McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press.

[24] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 311-318).

[25] Banerjee, S., & Lavie, A. (2005). METEOR. ACL Workshop on Evaluation Measures for MT, 65–72.

[26] Popovic, M. (2015). ChrF: Character n-gram F-score for automatic MT evaluation. Workshop on Statistical MT, 392–395.

[27] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157. https://doi.org/10.1007/BF02295996