Leveraging Data Triangulation Technique for the Development of Unwritten Language Parallel Corpora for Natural Language Processing Applications

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.
ahabuakar@abu.edu.ng

A. F. Donfack Kana
Department of Computer Science
Ahmadu Bello University
Zaria, Nigeria.
donfackkana@gmail.com

***Abstract*— Parallel corpora constitute the foundational datasets for contemporary neural machine translation (NMT) systems, with corpus quality and alignment accuracy directly impacting translation model performance. The scarcity of high-quality parallel corpora for low-resource Nigerian languages, compounded by their predominantly unwritten nature, presents significant barriers to NMT implementation. Current research only focus on developing parallel corpora with no consideration for unwritten languages or measures to validate and make the corpus reliable. This research addresses these challenges by developing a comprehensive English-Ebira parallel corpus utilizing dual-stage data triangulation methodologies and Text Similarity Matcher. The study implements both spatial and person-based triangulation techniques with a custom Text Similarity Matcher (TSM) program. The TSM incorporates Levenshtein distance and Jaro similarity metrics with optimized weighted scoring of (60 to 40 ratio) to optimize translation selection from multiple candidate translations to ensure corpus validity and reliability for the unwritten Ebira language corpus. The resulting corpus comprises 25,000 sentence pairs from religious texts, broadcast media, and adapted corpora. Experimental validation demonstrates 89.3% TSM accuracy and 8.8/10 overall quality validation rating by three individual Ebira language expert native speakers based on accuracy, fluency and cultural relevance of the corpus. This work contributes to advancing machine translation capabilities for unwritten languages and establishes a methodological framework applicable to similar low-resource language pairs.**

***Keywords*—unwritten languages, parallel corpora, data triangulation, text similarity, Ebira language, neural machine translation**

I. INTRODUCTION

The exponential growth in global digital communication has intensified demand for automated translation systems capable of bridging linguistic barriers efficiently and cost-effectively. Machine Translation (MT), a core application of computational linguistics and artificial intelligence, enables automated text translation between source and target languages through sophisticated algorithmic approaches [1]. Contemporary MT systems, particularly Neural Machine Translation (NMT) architectures, demonstrate superior performance compared to statistical and rule-based predecessors, yet require substantial parallel corpora for optimal functionality [2]. In African context, parallel corporal is a vital resource not just for building translation system, but also useful in retrieval of information across languages in the area of health, education and business transaction [3] and developing other NLP software like English-Ndebele medical dictionary [4].

Despite parallel corpora's critical importance for natural language processing (NLP) and machine translation research, most African languages, particularly Nigerian indigenous languages, remain underrepresented in digital linguistic resources [5]. This digital divide stems primarily from the unwritten nature of these languages and limited computational linguistics infrastructure.

Nigeria's linguistic landscape encompasses over 500 ethnic languages, representing extraordinary cultural and linguistic diversity. Among these languages, Ebira spoken by approximately two million native speakers across Kogi, Nassarawa, Federal Capital Territory, and Edo States exemplifies the challenges facing unwritten, low-resource languages in the digital age [6][7].

Unwritten languages, as defined by linguistic literature, are those lacking standardized orthographic systems or where written forms are not commonly utilized by native speaker communities [8]. Low-Resource Languages (LRLs) are characterized by insufficient digital content, limited computational tools, reduced scholarly attention, and minimal automated processing capabilities [9]. Such languages exhibit worldwide distribution spanning Africa, Pacific regions, Latin America, and Asia [10]. Notably, languages with established writing systems may still be classified as low-resource due to insufficient digital corpora required for NLP computational frameworks [11].

To address these challenges, this research employs data triangulation a systematic methodology involving multi-source data collection and cross-validation using TSM to ensure corpus reliability and validity [12]. The study's outcomes will facilitate NLP applications for Ebira language while establishing methodological benchmarks for other unwritten languages. Data acquisition encompasses diverse sources including English-Ebira religious texts, broadcast content from Tao FM and Jatto FM radio stations, digital resources from the Ebiradayi Worldwide organization, and adapted parallel corpora from existing Tatoeba datasets. All physical documents and PDF materials underwent digitization through Optical Character Recognition (OCR) technology to ensure computational accessibility and processing compatibility. This study aims to:

- Develop a high-quality English-Ebira parallel corpus using data triangulation methodologies
- Establish validation frameworks for ensuring corpus reliability in unwritten language contexts
- Create replicable methodologies for similar low-resource language corpus development
- Contribute to NLP resource development for Nigerian indigenous languages

II. RELATED WORK

A. African Language Parallel Corpora Development

Recent efforts in developing parallel corpora for African languages have targeted both high-resource languages (Hausa, Igbo, Yoruba, Swahili) and low-resource languages (Chichewa, Luganda, Igala). The WebCrawl African Corpora project covers 15 African languages with approximately 695,000 parallel sentences using automated web crawling from various internet sources [13]. Recent work has extended to unwritten languages like Tangale and Nupe through manual corpus creation methods [14].

B. Nigerian Language Focus

Nigerian language development has concentrated on three major languages: Hausa, Igbo, and Yoruba. The MENYO-20k English-Yoruba parallel corpus utilizes professional translation and manual verification methods [15]. For low-resource languages like Igala, researchers have employed both manual and machine-assisted translation approaches due to the unwritten nature of these languages.

C. Global Context and Methodology

Similar developments exist for Northeast Indian languages (Manipuri, Bodo, Mising, Karbi) [16] and Indic languages addressing script diversity and regional dialects [17]. European minority languages like Irish and Welsh have also benefited from parallel corpora development supporting language revitalization [18]. Corpus sizes for very low-resource languages typically range from 21,000 to 35,000 parallel sentences, with evaluation using BLEU and METEOR scores for machine translation quality assessment [19].

Also, String similarity algorithms have been extensively studied, with edit distance metrics like Levenshtein distance being among the most widely used [20]. The Levenshtein distance measures the minimum number of single-character edits required to transform one string into another [21]. While effective for capturing structural differences, it may be overly sensitive to minor orthographic variations common in low-resource languages.

Jaro similarity, designed for record linkage applications, demonstrates superior performance in handling character transpositions and minor spelling variations [22]. The Jaro-Winkler variant further improves performance by giving additional weight to common prefixes [23]. However, individual algorithms often fail to capture the full spectrum of similarity relationships in complex linguistic scenarios.

III. METHODOLOGY: DATA TRIANGULATION APPROACH

A. Dual-Stage Triangulation Framework

Data triangulation enhances reliability through multi-source data collection and cross-validation [24]. We implemented two complementary approaches:

Stage 1: Spatial Triangulation - Multi-source data acquisition from different geographical locations Stage 2: Person Triangulation - Independent translation by multiple translator groups

Stage 1: Multi-Source Data Acquisition

Religious Text Digitization: 7,000 sentence pairs extracted from English-Ebira Quran (5,000) and Bible (2,000) translations using OCR technology.

Lexical Resources: 3,114 sentence pairs from dictionary applications and web-based resources.

Broadcast Media: 1,886 parallel sentences from Jatto FM and Tao FM radio stations representing contemporary usage.

Corpus Adaptation: 13,000 English sentences adapted from Arabic-English and Dutch-English Tatoeba corpora for

translation. Table 1 below shows the summary of the corpus composition.

TABLE 1 CORPUS COMPOSITION

Source	Pairs	%	Characteristics
Religious texts	7,000	28%	Formal register, spiritual vocabulary
Lexical resources	3,114	12.5%	Basic vocabulary, everyday expressions
Broadcast media	1,886	7.5%	Contemporary usage, news terminology
Adapted corpora	13,000	52%	Diverse domains, varied complexity
Total	**25,000**	**100%**	**Comprehensive coverage**

B. Stage 2: Person Triangulation and TSM Processing

Three independent groups of professional Ebira native speakers translated the complete English corpus, producing trilingual Ebira translations as shown in Table 2. The Text Similarity Matcher system then selected optimal translations through weighted similarity analysis.

TABLE 2 SAMPLE TRANSLATION COMPARISON

English	Translator A	Translator B	Translator C	TSM Selection
Good morning	Nyene	Anyone	nyene	nyene
How are your people?	azauùyo?	Azawuyoo	azawu yoo?	azawuyoo
I'm happy for you	maje yiwu	Maajewu	majeyiwu	maje yiwu

C. Text Similarity Matcher (TSM) System

Architecture and Mathematical Foundation: The TSM implements hybrid similarity measurement designed for unwritten language characteristics with five core components: Text Normalization Engine, Multi-Algorithm Similarity Calculator, Decision Engine, Batch Processing Interface, and Quality Assurance Module. Equation (1) shows the primary similarity function of the TSM, while Equation (2) shows the equation for the calculation of text similarity weight with combined Levenstein and Jaro algorithm.

Primary Similarity Function:

$$S(A,B,C) = argmax\ \{S\_\{A,B\}, S\_\{A,C\}, S\_\{B,C\}\} \quad (1)$$

Weighted Similarity Calculation:

$$S\_\{X,Y\} = 0.6 * S\_\{Levenstein\}(X,Y) + 0.4 * S\_\{Jaro\}(X,Y) \quad (2)$$

Levenshtein Similarity: Handles structural variations and morphological changes Jaro Similarity: Addresses character transpositions and minor spelling variations

The 60:40 weighting was empirically optimized for Ebira language characteristics including vowel lengthening, consonant alternations, and morphological variations.

Algorithm Rationale: The weighted combination addresses specific Ebira text processing challenges:

- Vowel lengthening variations (e.g., "azauùyo" vs "Azawuyoo")
- Consonant alternations (e.g., "waadahi" vs "avudahi")
- Morphological variations (e.g., "maje yiwu" vs "majeyiwu")
- Orthographic inconsistencies due to lack of standardization

IV. RESULTS AND DISCUSSION

A. Overall Performance Analysis

TSM achieved 89.3% accuracy, representing a 7.7 percentage point improvement over the best single algorithm (Jaro at 81.6%) 90.1 accuracy. As shown in Table 3 below. Equation (3) and (4) were used to compute both accuracy and precision of TSM.

TSM Accuracy Calculation

$$\text{Accuracy} = \frac{\text{Number of Correct TSM Selections}}{\text{Total Number of Translation Triplets}} \quad (3)$$

Where:

- Correct TSM Selection: The TSM-selected translation matches the expert judgment
- Total Translation Triplets: All sets of three translations (A, B, C) processed

TSM Precision Calculation

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (4)$$

Where:

- True Positives (TP): TSM correctly identifies the best translation when one exists
- False Positives (FP): TSM selects a translation as "best" when it's not optimal

TABLE 3 TSM PERFORMANCE COMPARISON

Algorithm	**Accuracy (%)**	**Precision (%)**
Levenshtein Only	73.2	74.1
Jaro Only	81.6	82.3
TSM (Weighted)	**89.3**	**90.1**

B. Performance by Linguistic Phenomena

Highest Performance: Punctuation variations (94.8%) and vowel lengthening (92.1%) Most Challenging: Dialectal differences (86.3%) indicating need for enhanced regional variation handling as shown in Table 4.

TABLE 4: PERFORMANCE BY VARIATION TYPE

Variation Type	**Accuracy (%)**	**Key Examples**
Punctuation Variations	94.8	Question marks, exclamations
Vowel Lengthening	92.1	"azauùyo" → "Azawuyoo"
Consonant Variations	88.2	"waadahi" → "avudahi"
Morphological	87.9	"maje yiwu" → "majeyiwu"

C. Detailed Case Analysis

Example 1: Vowel Lengthening Recognition

English: "How are your people?"

A: "azauùyo?", B: "Azawuyoo", C: "azawu yoo?"

Similarity Scores generated using equation (2) above:

A to B: 0.847 (highest),

A to C: 0.692,

B to C: 0.731

Selection: "Azawuyoo" (standardized form) Correct

Example 2: Morphological Compound Handling

English: "I'm happy for you"

A: "maje yiwu", B: "Maajewu", C: "majeyiwu"

Similarity Scores:

A to B: 0.723,

A to C: 0.891 (highest),

B to C: 0.756

Selection: "maje yiwu" (canonical structure) Correct

D. Quality Assessment and Validation

The validation and assessment of the TSM system were carried out by combined three expert translators.

Corpus Statistical Characteristics:

- Total sentences: 25,000 English-Ebira pairs
- Average sentence length: 12.5 words (English), 11.9 words (Ebira)
- Unique vocabulary: 18,750 (English), 21,340 (Ebira)
- Translation consistency: 90.1% overall quality score

Inter-Translator Agreement:

- Perfect consensus (3/3): 23.4% occurrence, 100% TSM accuracy
- Majority agreement (2/3): 58.7% occurrence, 92.1% TSM accuracy
- No agreement (0/3): 17.9% occurrence, 81.3% TSM accuracy

Expert Validation Results: Independent validation of the parallel corpora by three Ebira language experts based on accuracy, fluency, cultural appropriateness and orthographic consistency of TSM yielded 8.8/10 overall quality rating on a likert scale of 1 to 10. The 8.8 is an average score of the three evaluators' 9, 8.6 and 8.8 individual overall ratings of the new English-Ebira corpora.

Performance Optimization

The weighted combination (60% Levenshtein, 40% Jaro) was optimized through grid search on a validation subset. As shown in Table 5 below, the optimal 60-40 weighting reflects the importance of structural similarity (Levenshtein) while maintaining sensitivity to character-level variations (Jaro) characteristic of Ebira texts.

TABLE 5 WEIGHT OPTIMIZATION RESULTS

Levenshtein Weight	Jaro Weight	Accuracy (%)
0.7	0.3	87.8
0.6	0.4	89.1
0.5	0.5	87.3

V. IMPACT AND APPLICATIONS

A. Community Impact and Future Research Directions

Engagement with Ebira speaking communities revealed significant positive impact of the TSM with the hope of having a true translation system for the Ebira language.

Corpus Expansion Plans:

- Target: 100,000 sentence pairs within 3 years
- Multimedia integration: Audio-text alignment for pronunciation resources

Technical Enhancements:

- Semantic similarity integration using word embeddings
- Deep learning applications with transformer-based models
- Real-time processing for mobile applications

VI. CONCLUSION

This research successfully demonstrates the feasibility of developing high-quality parallel corpora for unwritten, low-resource languages through systematic data triangulation.

The TSM system's weighted multi-algorithm approach effectively addresses orthographic variations characteristic of unwritten languages. The methodology demonstrates transferability across related languages and establishes frameworks for similar low-resource language initiatives.

Limitations include: scale constraints of manual translation, and dialectal representation gaps. Future work will focus on neural machine translation model development, corpus expansion, and semantic enhancement integration.

This work advances digital inclusion for African indigenous languages while providing methodological foundations for global endangered language preservation efforts. The English-Ebira corpus represents a significant step toward bridging the digital divide for low-resource African languages.

REFERENCES

[1] S. F. Ayegba, A. Onoja, and M. Ugbedeojo (2017). Igala parallel corpora for natural language processing applications. International Journal of Computer Applications, 171(9), 1–6.

[2] M. Ashraf. (2024). Innovation and Challenges in Neural Machine Translation: A Review, International journal of Science and Research (IJSR). https://doi.org/10.21275/SR241008113336

[3] K. Ogueji, D. I. Adelani, and I. Abdulmumin,(2024). NaijaNLP: A survey of Nigerian low-resource languages. arXiv preprint arXiv:2401.00000. https://arxiv.org/abs/2401.00000

[4] K. Ndhlovu (2016). Using ParaConc to extract bilingual terminology from parallel corpora: A case of English and Ndebele. Literator, 37(2), 1–11. https://doi.org/10.4102/lit.v37i2.1281

[5] G. George, O. Adebayo and C. Ogbodo (2024). TANGALENLP: Building Po Tangle to English parallel corpora and machine translation of the Tangale language. Proceedings of the 5th Workshop on African Natural Language Processing (AfricaNLP) at ICLR 2024. https://arxiv.org/abs/2404.07338

[6] J. Adive (1989). The verbal piece in Ebira. Dallas, TX: Summer Institute of Linguistics and the University of Texas at Arlington.

[7] Y. O. Aliyu (1989). In Preservation of An Identity: EPA lecture series. Kaduna: Nagazi printing press

[8] U.S. Election Assistance Commission. (2022, February). Best practices: Unwritten languages (Version 1.0). https://www.eac.gov/sites/default/files/Language_Access/Best_Practices_Unwritten_Languages_Final_508.pdf

[9] D. Robert (2014). Review of Developing Orthographies for Unwritten Languages [Review of the book Developing Orthographies for Unwritten Languages, by M. Cahill and K. Rice (Eds.)] Language Documentation and conservation, 8 781-788. http:hdl.handle.net/10125/24625

[10] Natural language processing applications for low-resource languages. (2025). Natural Language Processing. Cambridge University Press. https://www.cambridge.org/core/journals/natural-language-processing/article/natural-language-processing-applications-for-lowresource-languages/7D3DA31DB6C01B13C6B1F698D4495951

[11] J. L. García and D.C. Suárez (2022). Principles, scope, and limitations of the methodological triangulation. Journal of the Egyptian Public Health Association, 97, Article 48. https://pmc.ncbi.nlm.nih.gov/articles/PMC9714985/

[12] K. Ogueji, O. David, and T. Adewumi (2022). WebCrawl African Corpora: A multilingual parallel corpora for African languages. Proceedings of the 13th Language Resources and Evaluation Conference (LREC 2022), 1097–1107. https://aclanthology.org/2022.lrec-1.116/

[13] G. George, O. Adebayo and C. Ogbodo (2024). TANGALENLP: Building Po Tangle to English parallel corpora and machine translation of the Tangale language. Proceedings of the 5th Workshop on African Natural Language Processing (AfricaNLP) at ICLR 2024. https://arxiv.org/abs/2404.07338

[14] D. I. Adelani, J. Alabi, T. Adewumi, D. Ruiter, and S. Muhammad (2022). MENYO-20k: A multi-domain English–Yoruba parallel corpus for machine translation. Proceedings of the Language Resources and Evaluation Conference (LREC 2022), 1069–1076. https://aclanthology.org/2022.lrec-1.113/

[15] L. T. Atnafu, M. Mersha, A. Kalita, O. Kolesnikova and J. Kalita (2023). First attempt at building parallel corpora for machine translation of Northeast India's very low-resource languages. arXiv

[16] Anonymous. (2025). Parallel corpora for machine translation in low-resource Indic languages: A comprehensive review. arXiv.

[17] R. Doval. and A. Nieto. (2020). Using parallel corpora in the translation classroom. ERIC.

[18] S. P. Botley, A. M. McEnery, and A. Wilson. (2025). Applications of parallel corpora to the development of monolingual and parallel corpora for minority languages. CiteSeerX.

[19] V. I. Levenshtein. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, 10(8), 707-710.

[20] R. A. Wagner. and M. J. Fischer. (1974). The string-to-string correction problem. Journal of the ACM, 21(1), 168-173.

[21] M. A. Jaro (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. Journal of the American Statistical Association, 84(406), 414-420.

[22] W. E. Winkler. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. Proceedings of the Section on Survey Research Methods, 354-359.

[23] J. L. García, and D. C. Suárez (2022). Principles, scope, and limitations of methodological triangulation. Journal of Egyptian Public Health Association, 97.