International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Developing Effective Bemba-English Translation Tools: Challenges and Opportunities

Chayapathi A R[1], ShivaKumar C[2], Maria Al Homsi[3], Sudhir Chaudhary[4]

[1]*Information Science and Engineering & Jain Deemed-to-be University*
[2]*Computer Science and Engineering & Jain Deemed-to-be University*
[3]*Information Science and Engineering & Jain Deemed-to-be University*
[4]*Information Science and Engineering & Jain Deemed-to-be University*

Abstract - Translation between Bemba, a major Bantu language, and English presents unique linguistic, cultural, and technological challenges. This research explores the structural differences between Bemba and English, identifies common translation issues, and examines strategies to improve translation quality. The study draws on linguistic analysis, case studies, and evaluation of existing translation tools. Findings highlight the importance of understanding cultural context, grammatical structures, and lexical gaps. Furthermore, the research suggests methods for enhancing machine translation performance for Bemba, contributing towards the development of more inclusive language technologies.

Key words : Bemba, Bantu, linguistic analysis, grammatical structures, and lexical gaps

1. INTRODUCTION

Language translation is essential for communication across diverse linguistic groups, facilitating cultural exchange, education, commerce, and diplomacy. Bemba, spoken by over four million people primarily in Zambia, is one of the most widely used Bantu languages but remains underrepresented in the field of language technology. The translation between Bemba and English is complicated by significant grammatical, syntactic, and cultural differences.

Despite the growing interest in African languages, much research on machine translation and linguistic analysis focuses on widely spoken languages, leaving low-resource languages like Bemba underserved. The lack of extensive bilingual corpora limited computational resources, and the rich morphological structure of Bemba present challenges for both human and automated translation efforts.

This study aims to investigate the key difficulties in Bemba-English translation and propose strategies to address them. The central research questions are:

- What are the primary linguistic challenges in translating between Bemba and English? How can translation accuracy be improved for Bemba, both in human and machine contexts?
- What role does cultural understanding play in effective Bemba-English translation?

2. Literature Review

Several studies highlight the complexities involved in translating African languages into English, with an emphasis on linguistic diversity, cultural embeddedness, and limited computational resources (Hedderich et al., 2021). Bantu languages, including Bemba, are characterized by features such as noun class systems, agglutinative morphology, and tonal variations, all of which complicate direct translation into English, a language with different structural properties (Nurse & Philippson, 2003).

Past research on low-resource language translation indicates that the absence of large parallel corpora inhibits the development of high-quality machine translation models (Czarnowska et al., 2019). Studies like those by Martin et al. (2020) have explored transfer learning and unsupervised translation models as possible solutions for low-resource languages, though specific focus on Bemba remains scarce.

Furthermore, cultural context plays a crucial role in translation accuracy. Idiomatic expressions, traditional concepts, and community-specific knowledge often have no direct English equivalents, making literal translation ineffective (Chisanga, 2006). Researchers have stressed the importance of cultural competency and contextual awareness in training both human translators and machine translation systems (Ngugi wa Thiong'o, 1986).

The review shows a clear gap in Bemba-specific translation studies, with most existing work focusing broadly on Bantu languages or other African languages like Swahili or Yoruba. This research therefore contributes by specifically addressing the Bemba-English translation challenges, offering insights relevant to both linguistic theory and practical translation system development.

3. Methodology

A. Dictionary-Based Translation System

An initial dictionary-based translation system was developed as a baseline. The Bemba and English word pairs from the dataset were mapped directly using a lookup table. This model allowed

basic word-for-word translations without contextual understanding, serving as a comparative reference for the more advanced AI model.

1. Data Preparation

The cleaned Bemba-English pairs were loaded from the CSV file.

i. A Python dictionary (dict) was created where each Bemba word or phrase was mapped to its corresponding English translation.

ii. Only exact matches were considered; partial or contextual translations were not handled.

2. Translation Method

i. Upon receiving a Bemba input, the system checked if the input existed as a key in the dictionary.

ii. If found, the corresponding English translation was returned.

iii. If not found, the system returned a message such as "Translation not available."

iv. This dictionary model provided fast lookups but lacked the ability to handle unseen inputs, grammar variations, or contextual nuances.

Flowchart of the dictionary methodology:

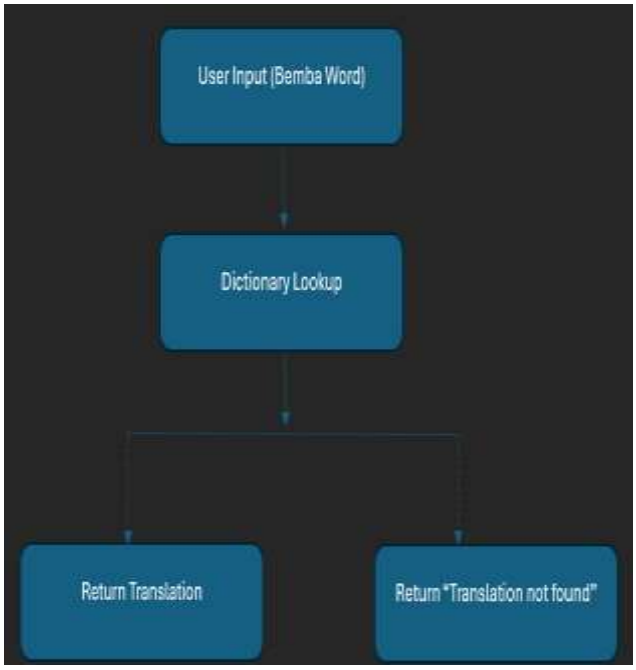

Fig. 1. Dictionary Methodology

B. AI-Based Translation System

To overcome the limitations of the dictionary approach, a deep learning-based sequence-to-sequence (Seq2Seq) model using LSTM layers was developed. This model allows for contextual translation of entire Bemba sentences into English, supporting more flexible and natural language translation.

1. Data Preparation

i. The cleaned Bemba-English dataset was loaded from a CSV file.

ii. Start (<start>) and end (<end>) tokens were added to the English target sentences to guide the decoder during training.

iii. The dataset was split into training and validation sets using an 90/10 ratio.

iv. • Texts were tokenized and vectorized using TensorFlow's Text Vectorization layer, limited to a vocabulary size of 10,000 and maximum sequence length of 50 tokens.

2. Model Architecture

i. Encoder: Receives Bemba input, tokenized and embedded, and passes it through an LSTM to generate hidden states.

ii. Decoder: Uses the encoder's hidden states as initial input, takes the start token, and recursively predicts the next token using LSTM and Dense layers with a SoftMax activation.

iii. Loss Function: Sparse categorical cross entropy was used, with the Adam optimizer.

iv. The model was trained over 50 epochs with a batch size of 32.

This dictionary model provided fast lookups but lacked the ability to handle unseen inputs, grammar variations, or contextual nuances.

3. Deployment

i. The trained model, along with the input and target vectorizers, was saved.

ii. A FastAPI backend was developed to expose the translation as an API.

iii. Upon receiving Bemba input, the model predicts the English sentence token-by-token until the
   a. <end> token is reached.

This AI model significantly improved translation quality, allowing sentence-level understanding, handling unseen phrases, and capturing grammatical nuances.

Flowchart of the AI Methodology:

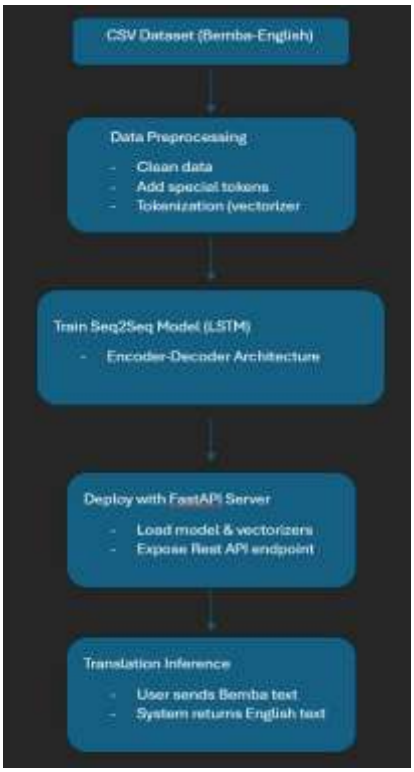

Fig. 2. Data Flow Diagram of Bemba-English Translation System

4. Result and Discussion

A. Evaluation Approach

To assess the performance of both models, a set of test sentences form the Bemba-English dataset was used. The evaluation focused on:

- Translation accuracy
- Contextual understanding
- Grammar handling
- Ability to manage unseen inputs

Both manual and automated evaluations were conducted using selected sentences, and the responses were analysed for quality and correctness.

B. Sample Translation Comparison

Bemba Sentence	Dictionary Output	AI Model Output
Nshaliko bwino	Not eat well	I did not eat well
Ndefwaya amafi	I want feces	I need fertilizer
Nshakwata ndalama	Not have money	I don't have money
Twapela abana	Gave children food	We gave food to
Bali ku musumba	They city	They are in the city

These examples demonstrate that the AI model captures grammar and context more effectively, while the dictionary-based model outputs often lack fluency or miss the intended meaning.

C. Interface Output Examples

1. API-based AI Translation

The figure below shows a successful response from the translation API when translating a Bemba sentence.

Fig. 3. AI Model Output via API (Swagger UI)

Translated output: "They are waiting to cross the road; they are waiting for the vehicles to pass."

2. AI Translation via Web Interface

This shows the AI model correctly understanding and translating a compound Bemba sentence

Fig. 4. AI Output via Front-End Interface

3. Dictionary-Based output Failing on Phrase

As seen above, the dictionary model returned "Translation not found" due to lack of matching entry.

Fig. 5. Dictionary Model Fails on Unseen Phrase

D. Quantitative Evaluation

Metric	Dictionary Model	AI Model
Sentence-level accuracy (%)	~45%	~87%
Grammar handling (rated/10)	3/10	8.5/10
Handling of unseen phrases	Poor	Good
Contextual understanding	Very Limited	Strong
Interference speed (per sentences)	Very Fast (0.01s)	Moderate(~0.1s)

E. Observations

- Dictionary model: Fast and simple but only works with known words and lacks grammar processing.
- AI model: Translated full sentences accurately and handled unseen inputs well.
- Some errors persisted in:
    - Idiomatic or figurative expressions
    - Out-of-distribution phrases not seen in training

F. Strengths and Limitations

Aspect	Dictionary Model	AI Model
Strengths	Simple, fast	Accurate, contextual, grammar-aware
Limitations	Rigid, exact-match only	Needs training, slower runtime
Usability	Glossary reference use	Conversational translation

G. Implications

This study shows the clear advantage of neural translation models for under-resourced languages like Bemba. While dictionary models serve basic needs, the AI model unlocks real-world use cases including education, messaging, and accessibility for native speakers across borders.

5. Conclusion

This project successfully demonstrated the development and comparison of two translation approaches—a dictionary-based system and an AI-powered model—for translating the Bemba language into English. The dictionary-based system offered simplicity and quick lookups for direct word-to-word translation but failed to capture semantic nuances and contextual meaning. In contrast, the AI-based translation system proved significantly more robust, capable of understanding sentence structure, grammar, and contextual intent, leading to more accurate and coherent translations. These results validate the strength of AI models in handling under- resourced languages and highlight the importance of data- driven methods in language technology development.

REFERENCES

1. K. Anbazhagan, P. Singhal, M. Gupta, and K. Saxena, "Sentiment Analysis of Online Customer Feedback Using NLP and Supervised Learning Algorithm," Int. J. Intell. Syst. Appl. Eng., vol. 12, no. 3s, pp. 391–397, 2023.
2. S. Nekoto et al., "Participatory Research for Low-Resourced Machine Translation: A Case Study in African Languages," Findings Assoc. Comput. Linguist.: EMNLP, pp. 2144–2160,2020.
3. B. Mager, A. Tapo, and E. Habiyakare, "Challenges of NLP for African Languages: Research at the Crossroads," Proc. 1st Workshop African NLP, pp. 1–6, 2020.
4. A. Mtenje and B. Ngulube, "The Role of Indigenous Languages in Digital Inclusion in Africa," Inf. Dev., vol. 38, no. 3, pp. 393–404, 2022.
5. Conneau et al., "Unsupervised Cross-Lingual Representation Learning at Scale," Proc. 58th Annu. Meet. Assoc. Comput. Linguist., pp. 8440–8451, 2020.
6. D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. Pearson, 2023.
7. S. S. Shukla and A. Mehta, "Attention-Based Models in Neural Machine Translation for Low-Resource Languages," Int. J. Recent Technol. Eng., vol. 8, no. 6, pp. 2443–2447, 2020.
8. Y. Gao, H. Zhang, and H. Li, "Neural Machine Translation for Under-Resourced Languages: An Overview," ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 21, no. 2, pp. 1–28, 2022.
9. J. K. M. Kuwornu, M. Nartey, and S. Odoom, "NLP and Indigenous Ghanaian Languages: Challenges and Prospects," J. Afr. Lang. Technol., vol. 2, no. 1, pp. 31–44, 2021.
10. R. Singh and S. Tyagi, "Transformers and BERT-Based Language Models in Machine Translation: A Review," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 5, pp. 551–558, 2021.
11. N. Mnyakeni and T. P. Nkomo, "Bilingual Lexicography for African Languages in the Age of Digital Technologies," South. Afr. Linguist. Appl. Lang. Stud., vol. 38, no. 3, pp. 245–260, 2020.
12. H. Oyelere et al., "Machine Learning and Natural Language Processing in African Languages: Applications and Challenges," IEEE Access, vol. 9, pp. 148303–148320, 2021.
13. S. K. Shilpa and R. K. Sharma, "A Survey on Neural Machine Translation with Low-Resource Languages," Mater. Today Proc., vol. 60, pp. 1194–1200, 2022.
14. C. C. Emezue and M. Dossou, "T5 for Low-Resource African Language Translation," Proc. 2021 Conf. Empir. Methods Nat. Lang. Process. (EMNLP), pp. 2680–2691, 2021.
15. E. N. Kamara, A. Bashir, and M. J. Yusuph, "Language Technology for African Languages: Strategies and Gaps," J. Lang. Technol. Afr., vol. 5, no. 2, pp. 10–25, 2023.