

Resource Sufficient Data Generation Technique for Machine Translation Models Using Named Entity Transliteration and Cross- Lingual Word Embeddings

*¹Ibrahim Rabi Abdullahi, ²Georgina Nkolika Obunadike, ³Adam Abdullahi and ⁴Martins E. Irhebhude

¹Division of Agricultural Colleges, Ahmadu Bello University, Zaria, Kaduna State, Nigeria.

²Federal University, Dutsin-Ma, Katsina State, Nigeria.

³Kaduna State University, Kaduna State, Nigeria.

⁴Computer Science Department, Nigerian Defence Academy, Kaduna, Nigeria.

*Corresponding authors' email: iarabiu@abu.edu.ng

ABSTRACT

Machine translation offers speed and cost advantages, but human translation remains superior in quality and contextual understanding. For low-resource languages like Hausa, the severe scarcity of parallel data leads to poor model performance, overfitting, and inadequate handling of named entities. While existing synthetic data generation methods such as back-translation and data copying introduce noise, this study proposes a technique that integrates named entity transliteration with cross-lingual word embeddings to generate synthetic parallel sentences. The approach replaces the common practice of entity copying with transliteration to preserve target-language vocabulary integrity, while embeddings translate non-entity words. Evaluated on the FLORES-200 and MAFAND-MT benchmarks, the unsupervised setting achieved Bilingual Evaluation Understudy (BLEU) scores of 1.04 (Hausa→English) and 1.42 (English→Hausa) on FLORES-200, while the semi-supervised setting improved substantially to 15.56 (Hausa→English) and 12.42 (English→Hausa) BLEU on MAFAND-MT outperforming the monolingual data copying baseline by up to 5 BLEU points. These results demonstrate that the proposed technique offers an effective strategy for augmenting scarce parallel data in low-resource MT systems.

Received: 30 Jul. 2026

Accepted: 03 Oct. 2026

Published: 06 Oct. 2026

Keywords:

Natural language processing, Low-resource language, Hausa-English, Machine Translation

INTRODUCTION

Effective communication fosters understanding and collaboration across cultures, bridging language gaps between under-resourced languages like Hausa and their counterparts, empowering individuals to participate actively in social, economic, and educational opportunities (Zahorodna et al., 2022). Machine Translation (MT) systems provide key advantages in terms of speed, affordability, and accessibility, while human translation remains superior in quality, detail, and contextual awareness (Siu, 2023). However, the performance of MT systems depends on the availability of large-scale, high-quality parallel corpora aligned sentence pairs in source and target languages (Raja & Vats, 2025).

For high-resource language pairs such as English–French or English–Chinese, massive parallel datasets enable the training of state-of-the-art neural machine translation (NMT) models that achieve near-human performance (Skenduli et al., 2023). In contrast, low-resource languages (LRLs) suffer from a severe scarcity of parallel data, a problem identified as a primary barrier to developing effective MT systems for these languages (Mager et al., 2023; Tafa et al., 2025).

Hausa, a Chadic language spoken by over 80 million people primarily in Nigeria, Niger, and Ghana, exemplifies this challenge (Muhammad et al., 2025). Despite its large speaker population, Hausa remains severely under-resourced in terms of digital language resources, parallel corpora, and NLP tools (Adelani et al., 2021; Olojo et al., 2025). Manual creation of parallel corpora is prohibitively expensive and time-consuming, particularly for language pairs where professional translators are scarce (Thai et al., 2022).

Consequently, researchers have increasingly turned to automated data generation methods to augment limited parallel resources. However, existing methods have critical limitations: back-translation introduces noise (Burchell et al., 2022), data copying risks polluting target vocabulary (Lee et al., 2022), and word-level translation struggles with out-of-vocabulary items (Ibrahim & Abdulmumin, 2022).

A particular challenge in low-resource MT is the handling of named entities. Named Entity Recognition (NER) the automatic identification and classification of textual references into persons, locations, and organizations (Li et al., 2022) is crucial for translation accuracy. While initiatives such as MasakhaNER have provided benchmarks for African languages including Hausa (Adelani et al., 2021; Ojo et al., 2023), small datasets continue to limit performance. A related challenge is named entity transliteration, which preserves phonetic form when converting orthographic representations across scripts (Babayev, 2025). Cross-lingual word embeddings capture semantic and syntactic similarities between words across languages, enabling accurate translations even for lexically unmapped words (Sharma et al., 2023). While cross-lingual embeddings have been explored for various tasks, their combination with named entity transliteration for data generation remains unexplored. This study addresses these gaps through a novel data generation technique for English–Hausa MT that integrates named entity transliteration with cross-lingual word embeddings. Specifically, this study makes three primary contributions:

- i. A novel data generation methodology that integrates named entity transliteration with cross-lingual word embeddings for English – Hausa MT.
- ii. An empirical demonstration that the proposed approach outperforms existing synthetic data methods in semi-supervised settings.
- iii. A publicly available transliteration databank for English–Hausa to support future research.

Our technique generates synthetic parallel sentences that augment scarce resources for training NMT models. We evaluate our approach on both unsupervised and semi-supervised settings, using the FLORES-200 (Goyal et al., 2022) and Mafand-MT (Adelani et al., 2022) benchmarks. To the best of our knowledge, this is the first study to combine named entity transliteration and cross-lingual word embeddings specifically for English–Hausa MT data generation.

The remainder of this paper is organized as follows. Section 2 describes the materials and methods, including datasets, the proposed algorithm, and experimental setup. Section 3 presents the results and discussion. Section 4 concludes with recommendations for future work.

MATERIALS AND METHODS

This section describes the datasets, pre-processing procedures, pre-trained models, and the proposed algorithm for generating synthetic parallel data for English–Hausa machine translation.

Datasets

Four primary data sources were used in this study:

MAFAND-MT

A parallel corpus for African languages containing news and web domain sentences (Adelani et al., 2022). This served as the seed dataset for semi-supervised training.

MasakhaNER 2.0

A named entity recognition benchmark for 21 African languages, including Hausa (Ojo et al., 2023). This was used for NER extraction.

PanLex Bilingual Dictionary

A lexical resource covering over 5,700 languages, used for supervised cross-lingual embedding alignment (Mager et al., 2023).

Monolingual Corpora

Hausa and English sentences scraped from Premium Times news websites, used as source text for synthetic data generation.

Data Pre-processing

We used MAFAND-MT as our primary parallel corpus, which required little pre-processing since it is already sentence-aligned. In contrast, the PanLex dictionary posed a challenge: it contains accented Hausa terms (e.g., 'gádo' rather than 'gado') that our embedding models cannot process. Normalizing these accents yielded 7,139 usable word pairs—a modest but valuable resource for aligning English and Hausa embeddings.

Monolingual sentences scraped from web pages underwent additional pre-processing: removal of URLs, HTML elements, lines consisting entirely of letters without words, and lines containing fewer than five words.

Pre-trained Models

Word Embeddings

Two Word2Vec models were employed: word2vec-google-news-300 (3 million English word vectors) and HauWE (90,454 Hausa word vectors), both with 300-dimensional vectors (Muhammad et al., 2021). The English embedding was reduced from 10.03 GB to approximately 331 MB by selecting the 100,000 most frequent terms.

Using the VecMap framework, supervised and unsupervised cross-lingual word embeddings were trained to enable word-level translation between Hausa and English. The supervised method leveraged the PanLex bilingual dictionary, while the unsupervised method relied solely on monolingual data.

Named Entity Recognition

Named entities were extracted using two tools:

- i. MasakhaNER 2.0 – Fine-tuned on 21 African languages, supporting PERSON, LOCATION, ORGANIZATION, and DATE extraction (Ojo et al., 2023).
- ii. NLTK Named-Entity Tagger (default NLTK classifier) – Used for English text, additionally recognizing TIME and MONEY entities (Au et al., 2022).

Machine Translation Model

We fine-tuned the M2M-100 multilingual translation model (Fan et al., 2021), which supports direct translation between 100 languages without English-centric bias. A pre-trained checkpoint was obtained from HuggingFace and fine-tuned on MAFAND-MT and synthetically generated data.

Synthetic Data Generation Algorithm

Algorithm 1 outlines the proposed method for generating synthetic parallel sentences, illustrated in Figure 1.

Algorithm 1: Synthetic Translation Generation (Hausa↔English)

Input: Monolingual Sentence, CLWE, NE and Transliteration Dictionary

- 1: Tokenization (retaining word indices)
- 2: Extract Named Entities
- 3: Extract Non-Named Entities
- 4: Use CLWE to translate Non-Named Entities. If they exist, else reuse the word
- 5: Use a transliteration dictionary to replace Named Entities. If they exist, else reuse the original word (token)
- 6: Use indices in (1) to build the translation using generated words in (4) and (5).

Output: Synthetic Translation

A transliteration dictionary was manually compiled from online Hausa news sources and hardcopy newspapers,

containing 1,500 named entity pairs for English–Hausa transliteration.

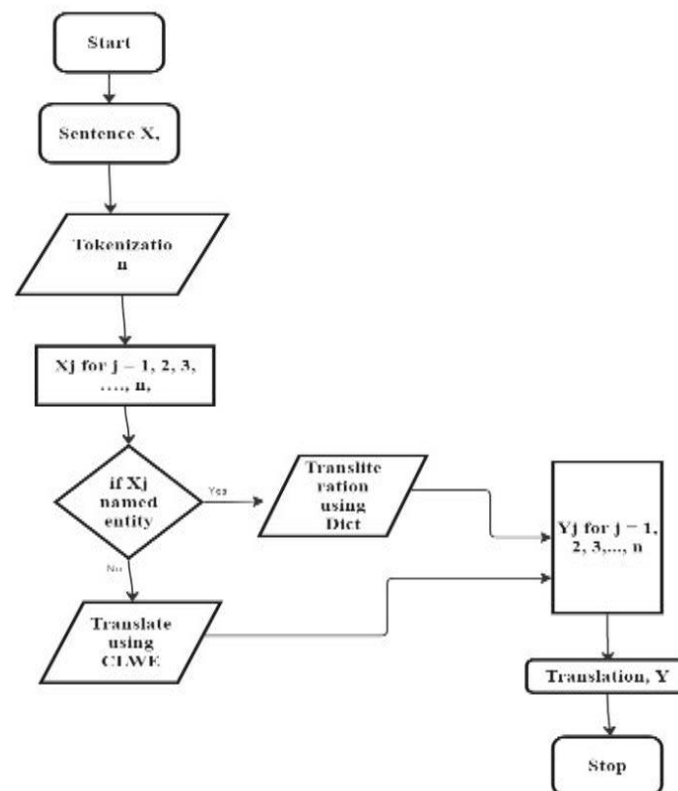


Figure 1: chart of the data creation process

Experimental Setup

Training Data

The MAFAND-MT training split (5,865 sentence pairs) served as the seed corpus for semi-supervised learning. Monolingual Hausa and English sentences crawled from Premium Times were used as source data for synthetic generation, yielding synthetic datasets of 2,556 (small) and 3,896 (large) sentences after pre-processing. For the semi-supervised setting, these were combined with the 5,865-sentence MAFAND-MT seed, producing training sets of 8,421 (30.4% synthetic) and 9,761 (39.9% synthetic) sentences respectively.

Training Settings

The M2M-100 model was fine-tuned under two settings:

- i. Unsupervised: No parallel data; relied on cross-lingual embedding alignment.
- ii. Semi-supervised: Used a small seed of MAFAND-MT parallel data augmented with synthetic translations.

Word embeddings were trained with a dimension of 300 using the Skip-gram architecture. The PanLex dictionary was used for supervised embedding alignment. The model was fine-tuned for 10 epochs with a batch size of 32, a learning rate of $5e-5$, and the Adam optimizer. The MAFAND-MT dataset was split into 80% training, 10% validation, and 10% test sets. The MAFAND-MT test set contains 1,024 sentences, while the FLORES-200 test set contains 1,012 sentences per language for the English–Hausa pair. Models were evaluated

on FLORES-200 and MAFAND-MT test sets using SacreBLEU (Post, 2018).

RESULTS AND DISCUSSION

This section presents the results of the word-level translation experiments and the machine translation performance of models trained on synthetic data generated using the proposed approach. All models were evaluated on the MAFAND-MT and FLORES-200 test sets using SacreBLEU.

Word-Level Translation with Cross-Lingual Embeddings

Figures 2a and 2b illustrate the performance of the word-level translation models using the VecMap framework with four matching algorithms: Nearest Neighbor (nn), Inverse Nearest Neighbor (inv-nn), cross-domain similarity local scaling (csls), and inverse softmax (inv-softmax). Translation quality was evaluated using a human-assigned grading scale of 0 to 3:

- i. Grade 0: Untranslated - source word not translated; may be absent from the CLWE vocabulary.
- ii. Grade 1: Poor translation - returned target word shares no semantic relationship with the source.
- iii. Grade 2: Good translation - returned target word shares some semantic relationship but is not an exact match.
- iv. Grade 3: Perfect translation - returned target word is an exact semantic match.

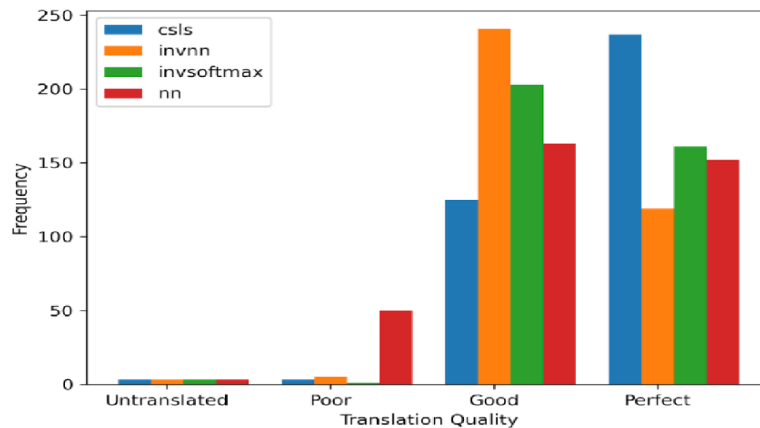


Figure 2a: Human evaluation of English to Hausa word-level translation quality.

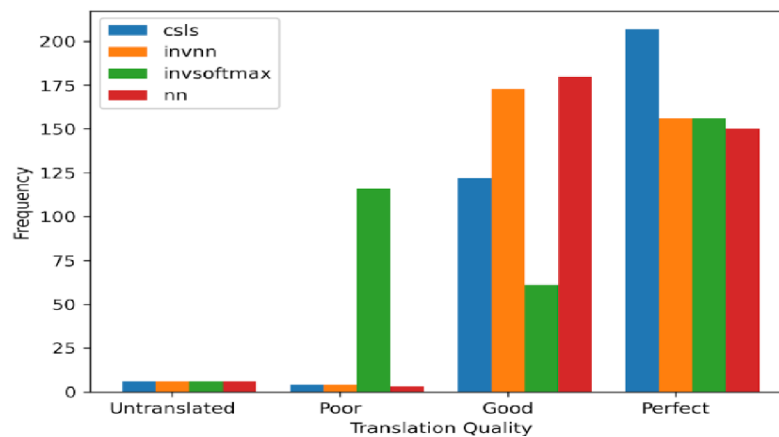


Figure 2b: Human evaluation of Hausa to English word-level translation quality.

Evaluation revealed six untranslated words from the Hausa test data and three from English, indicating vocabulary gaps in the cross-lingual embedding models. The CSLS algorithm produced the highest number of perfect translations and was therefore selected for generating the synthetic parallel data.

Machine Translation Experiments

The M2M-100 model was fine-tuned under two settings:

- Unsupervised: Trained solely on synthetic data generated using the proposed approach.
- Semi-supervised: Trained on a mixture of synthetic data and the MAFAND-MT human-translated seed corpus.

Synthetic datasets of varying sizes were generated:

- Small synthetic set: 2,556 sentences
- Large synthetic set: 3,896 sentences

For semi-supervised experiments, these were combined with a seed of 5,865 human-translated MAFAND-MT sentences, yielding training sets of 8,421 and 9,761 sentences respectively.

Performance of Proposed Approach (Transliteration)

Table 1 presents the BLEU scores for M2M-100 models trained on synthetic data generated using cross-lingual word embeddings with named entity transliteration.

Table 1: Performances of M2M-100 Models Trained with Synthetic Data Using Transliteration

Model	Setting	Data Size	Hausa → English		English → Hausa	
			(MAFAND-MT)	(FLORES-200)	(MAFAND-MT)	(FLORES-200)
M2M-100 fine-tuned	Unsupervised	2,556	0.1372	1.0415	0.8016	1.4217
M2M-100 fine-tuned	Unsupervised	3,896	0.1312	0.9129	0.7312	3.5521
M2M-100 fine-tuned	Semi-supervised	8,421	15.5598	9.8814	12.4221	9.4119
M2M-100 fine-tuned	Semi-supervised	9,761	15.1147	9.0197	12.4904	9.1913

In the unsupervised setting, models performed better on FLORES-200 than on MAFAND-MT. Smaller synthetic datasets yielded slightly better performance than larger ones, suggesting that lower-quality synthetic data degrades performance when added in larger quantities. In the semi-supervised setting, performance improved substantially—by over 15 BLEU points in some cases—and models performed better on MAFAND-MT than on FLORES-200. This is likely

due to domain similarity between the MAFAND-MT training and test sets. Nevertheless, performance on FLORES-200 improved by up to 5.9 BLEU points.

Comparison with Human-Translated Data

Table 2 compares M2M-100 models trained on human-translated MAFAND-MT data against models trained solely on synthetic data.

Table 2: M2M-100 performances on human-translated vs. synthetic data

Model	Training Data	Data Size	Hausa → English		English → Hausa	
			MAFAND-MT	(FLORES-200)	(MAFAND-MT)	(FLORES-200)
M2M-100 fine-tuned	MAFAND-MT only	5,865	16.0818	12.66798	12.5203	9.0838
M2M-100 fine-tuned	Synthetic only	3,896	2.44393	6.6833	0.4452	2.6264

Models trained solely on human-translated data outperformed those trained on synthetic data, as expected given the higher quality of human translations. However, the semi-supervised approach (Table 3.1) substantially closed this gap, achieving performance close to the human-translated baseline.

Comparison with Named Entity Replacement

Table 3 presents results for M2M-100 models trained on synthetic data generated using named entity replacement instead of transliteration. Values in parentheses indicate the difference from the transliteration approach (Table 3.1).

Table 3: M2M-100 performances with named entity replacement (difference from transliteration)

Model	Setting	Data Size	Hausa → English		English → Hausa	
			(MAFAND-MT)	(FLORES-200)	(MAFAND-MT)	(FLORES-200)
M2M-100 fine-tuned	Unsupervised	2,556	0.1373 (+0.001)	1.0142 (-0.027)	0.7953 (-0.006)	1.3419 (-0.080)
M2M-100 fine-tuned	Unsupervised	3,896	0.1295 (-0.002)	0.8707 (-0.042)	0.6940 (-0.037)	3.5930 (+0.041)
M2M-100 fine-tuned	Semi-supervised	8,421	14.6910 (-0.869)	9.5729 (-0.309)	12.3241 (-0.098)	8.6716 (-0.740)
M2M-100 fine-tuned	Semi-supervised	9,761	14.9182 (-0.197)	8.5675 (-0.452)	12.2334 (-0.257)	8.9761 (-0.215)

The transliteration approach consistently outperformed entity replacement, validating the superiority of the proposed method.

Comparison with Monolingual Data Copying

Table 4 compares the proposed approach with the monolingual data copying method of Currey et al. (2017) using M2M-100 models.

Table 4: M2M-100 Performances with Monolingual Data Copying (Currey et al., 2017)

Model	Setting	Data Size	Hausa → English		English → Hausa	
			(MAFAND-MT)	(FLORES-200)	(MAFAND-MT)	(FLORES-200)
M2M-100 fine-tuned	Unsupervised	2,556	0.2782	2.5810	0.2854	2.6127
M2M-100 fine-tuned	Unsupervised	3,896	0.2779	2.4940	0.2753	2.5015
M2M-100 fine-tuned	Semi-supervised	8,421	10.2781	7.5062	11.2807	8.5177
M2M-100 fine-tuned	Semi-supervised	9,761	10.1023	7.2681	11.6051	8.1679

In the unsupervised setting, data copying outperformed the proposed approach, possibly due to the disjointed grammatical structure of word-level translations. However, in the semi-supervised setting, both proposed approaches outperformed Currey et al. by up to 5 BLEU points. This is because the source and target sentences in the proposed approach are in different languages, whereas Currey's method copies source-language data, causing language confusion when mixed with authentic parallel data.

Discussion

The results demonstrate that while human-translated data remains superior, synthetic data generation using word-level translation with cross-lingual embeddings and named entity transliteration provides a viable strategy for augmenting scarce parallel resources. The semi-supervised approach, which combines a small seed of human translations with larger synthetic datasets, achieves performance approaching that of models trained solely on human data.

Several avenues for improvement exist. First, higher-quality cross-lingual embeddings would improve word-level translation accuracy. Second, a richer transliteration dictionary would enhance named entity handling. Third, periodic retraining of the embedding models with new data sources would expand vocabulary coverage and reduce untranslated words.

CONCLUSION

This study proposed a resource-sufficient data generation technique for low-resource machine translation using named entity transliteration and cross-lingual word embeddings for English–Hausa and Hausa–English translation.

The methodology involved: (1) compiling a transliteration databank of **1,500** named entity pairs from online Hausa news sources and hardcopy newspapers; (2) training cross-lingual word embeddings using the VecMap framework with both supervised (using the PanLex bilingual dictionary) and unsupervised approaches; (3) generating synthetic parallel sentences using Algorithm 1, which combines tokenization, named entity recognition, CLWE translation for non-entities, and transliteration for named entities; and (4) fine-tuning the M2M-100 multilingual translation model under unsupervised and semi-supervised settings.

The experimental results demonstrate the effectiveness of the proposed approach. In the unsupervised setting, the model achieved BLEU scores of **1.04** (Hausa→English) and 1.42 (English→Hausa) on the FLORES-200 test set, and 0.14 and 0.80 on the MAFAND-MT test set. The semi-supervised setting improved substantially, achieving BLEU scores of 15.56 (Hausa→English) and 12.42 (English→Hausa) on MAFAND-MT, and 9.88 and 9.41 on FLORES-200—representing an improvement of over 15 BLEU points in some cases compared to the unsupervised setting.

The proposed transliteration approach consistently outperformed named entity replacement, with improvements of up to 0.87 BLEU in semi-supervised settings. Furthermore, the proposed method outperformed the monolingual data copying baseline (Currey et al., 2017) by up to 5 BLEU points in semi-supervised settings, validating the superiority of transliteration over entity copying for preserving target-language vocabulary integrity.

The core problems of previous work—out-of-vocabulary words, limited vocabulary coverage, and lack of named entity transliteration—were addressed through the manually

compiled transliteration dictionary and cross-lingual embeddings. While human-translated data remains superior (16.08 BLEU on MAFAND-MT), the proposed approach provides a viable and cost-effective strategy for augmenting scarce parallel resources in low-resource language pairs.

Several avenues for future improvement exist. First, higher-quality cross-lingual embeddings would improve word-level translation accuracy. Second, expanding the transliteration dictionary beyond 1,500 pairs would enhance named entity handling. Third, periodic retraining of the embedding models with new data sources would expand vocabulary coverage and reduce untranslated words. This approach not only enhances translation model performance for English–Hausa and Hausa–English pairs but also provides a foundation for similar improvements in other low-resource language translations.

Data Availability Statement

The transliteration databank of 1,500 English–Hausa named entity pairs is available at https://github.com/abunimra/Rabiu_Msc.git. The MAFAND-MT and FLORES-200 datasets are publicly available from their respective sources. The code used for synthetic data generation and model fine-tuning is also available at the same repository.

REFERENCES

- Adelani, D. I., Abbott, J., Neubig, G., Daniel, D., Kreutzer, J., Lignos, C., Palen-michel, C., Buzaaba, H., Rijhwani, S., Ruder, S., Mayhew, S., & Azime, I. A. (2021). MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9, 1116–1131. https://doi.org/10.1162/tacl_a_00408
- Adelani, D. I., Alabi, J. O., Fan, A., Kreutzer, J., Shen, X., Reid, M., Ruiter, D., Klakow, D., Nabende, P., Chang, E., Gwadabe, T. R., Sackey, F., Dossou, B. F. P., Emezue, C. C., Leong, C., Beukman, M., Muhammad, S. H., Jarso, G. D., Yousuf, O., ... Manthalu, S. (2022). A few thousand translations go a long way! Leveraging pre-trained models for African news translation. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3053–3070. <https://doi.org/10.18653/v1/2022.naacl-main.223>
- Au, T. W. T., Lamos, V., & Cox, I. J. (2022). E-NER: An annotated named entity recognition corpus of legal text. *Proceedings of the Natural Legal Language Processing Workshop 2022*, 246–255. <https://doi.org/10.18653/v1/2022.nllp-1.22>
- Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Grave, E., Auli, M., & Joulin, A. (2021). Beyond English-centric multilingual machine translation. *Journal of Machine Learning Research*, 22, 1–48.
- Goyal, N., Gao, C., Chaudhary, V., Chen, P. J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., & Fan, A. (2022). The FLORES-200 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10, 522–538. https://doi.org/10.1162/tacl_a_00474
- Ibrahim, R. A., & Abdulmumin, I. (2022). NECAT-CLWE: A simple but efficient parallel data generation approach for low resource. *AfricaNLP Workshop at the International Conference on Learning Representations (ICLR) 2022*, 1–8.
- Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., & Carlini, N. (2022). Deduplicating training data makes language models better. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1, 8424–8445. <https://doi.org/10.18653/v1/2022.acl-long.577>
- Li, J., Sun, A., Han, J., & Li, C. (2022). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1), 50–70. <https://doi.org/10.1109/TKDE.2020.2981314>
- Mager, M., Bhatnagar, R., Neubig, G., Thang, N., & Katharina, V. (2023). Neural machine translation for the indigenous languages of the Americas. *Proceedings of the Association for Computational Linguistics*, 109–133.
- Muhammad, P. F., Kusumaningrum, R., & Wibowo, A. (2021). Sentiment analysis using Word2vec and long short-term memory (LSTM) for Indonesian hotel reviews. *Procedia Computer Science*, 179, 728–735. <https://doi.org/10.1016/j.procs.2021.01.061>
- Muhammad, S. H., Ahmad, I. S., Abdulmumin, I., Lawan, F. I., Imam, S. H., Aliyu, Y., Sani, S. A., Umar, A. U., Gwadabe, T., Church, K., & Marivate, V. (2025). HausaNLP: Current status, challenges and future directions for Hausa natural language processing. *Proceedings of the Sixth Workshop on African Natural Language Processing*, 176–191. <https://doi.org/10.18653/v1/2025.africanlp-1.27>
- Ojo, J., Ogueji, K., Stenetorp, P., & Ifeoluwa, D. (2023). How good are large language models on African languages? [Unpublished manuscript or conference paper].
- Olojo, S., Zakrzewski, J., Smart, A., Van Liemt, E., Miceli, M., Ebinama, A., & Amugongo, L. M. (2025). Lost in machine translation: The sociocultural implications of language technologies in Nigeria. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 1(1). <https://doi.org/10.1145/3715275.3732153>
- Post, M. (2018). A call for clarity in reporting BLEU scores. *Proceedings of the Third Conference on Machine Translation: Research Papers*, 186–191.
- Raja, R., & Vats, A. (2025). Parallel corpora for machine translation in low-resource Indic languages: A comprehensive review. *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages*, 129–143. <https://doi.org/10.18653/v1/2025.loresmt-1.12>
- Sharma, S., Diwakar, M., Singh, P., Singh, V., Kadry, S., & Kim, J. (2023). Machine translation systems based on classical-statistical-deep-learning approaches. *Electronics*, 12(7), 1–29. <https://doi.org/10.3390/electronics12071716>
- Siu, S. C. (2023). ChatGPT and GPT-4 for professional translators: Exploring the potential of large language models

- in translation. *SSRN Electronic Journal*, 1–36. <https://doi.org/10.2139/ssrn.4448091>
- Skenduli, M. P., Shekhar, R., Alam, M., & Kaur, R. (2023). Neural machine translation for low-resource languages. *ACM Computing Surveys*, 55(11), 1–37.
- Tafa, T. O., Hashim, S. Z. M., Othman, M. S., Alhussian, H., Nasser, M., Abdulkadir, S. J., Huspi, S. H., Adeyemo, S. O., & Bena, Y. A. (2025). Machine translation performance for low-resource languages: A systematic literature review. *IEEE Access*, 13, 72486–72505. <https://doi.org/10.1109/ACCESS.2025.3562918>
- Thai, K., Karpinska, M., Krishna, K., Ray, W., Inghilleri, M., Wieting, J., & Iyyer, M. (2022). Exploring document-level literary machine translation with parallel paragraphs from world literature. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9882–9902. <https://doi.org/10.18653/v1/2022.emnlp-main.672>
- Zahorodna, O., Saienko, V., Tolchieva, H., Tymoshchuk, N., Kulinich, T., & Shvets, N. (2022). Developing communicative professional competence in future economic specialists in the conditions of postmodernism. *Postmodern Openings*, 13(2), 77–96.

