A Low-Resource Language Translation: French To Mooré

Hamed Joseph Ouily, Aminata Sabané, Delwende Eliane Birba, Rodrique Kafando, Abdoul-Kader Kabore, Tégawendé F. Bissyandé

A Low-Resource Language Translation: French To Mooré

Hamed Joseph Ouily[*1,2], **Aminata Sabané**[1,2], **Delwende Eliane Birba**[3], **Rodrique Kafando**[1], **Abdoul-Kader Kabore**[1], **Tégawendé F. Bissyandé**[12]

[1]Centre d'Excellence CITADEL, Université Virtuelle du Burkina Faso
[2]Département d'Informatique, UFR/SEA, Université Joseph KI-ZERBO, Burkina Faso
[3]AI-KING GROUP

*E-mail : hamed.ouily@gmail.com

Abstract
Natural Language Processing (NLP) is an exciting field of artificial intelligence with the goal of enabling machines to understand human language in a natural way. Neural Machine Translation (NMT) stands out as one of the most promising applications of NLP, offering the ability to effectively translate text from a source language to a target language. In recent years, NMT has experienced significant advances, marking a major milestone in the development of automatic translation systems. Through the use of neural networks, NMT has demonstrated an ability to capture the nuances of language, thereby improving the quality of translations and making the experience of multilingual communication more seamless and precise. This evolution has opened new perspectives in areas such as international collaboration, intercultural understanding, and the global dissemination of information. However, most African languages, especially those in Burkina Faso, have received very little research attention in this context. In this article, we propose automated translation models *French* to *Mooré* based on Transformers. We achieved a BLEU score of 71.18 for the automated for the second model, *French* to *Mooré* translation.

Keywords
Natural Language Processing, Neural Machine Translation, Low-ressource Language, Local language, Mooré Language

I INTRODUCTION

Linguistic diversity is one of the riches of Africa [14]. Languages hold strategic importance for both peoples and the planet, as they play a crucial role in the development process. They represent the wealth of cultural diversity and facilitate intercultural dialogue. Additionally, languages are an essential tool for ensuring quality education accessible to all. They encourage collaboration and contribute to the establishment of inclusive knowledge societies. They also preserve precious cultural heritage and stimulate political commitment to the beneficial application of science and technology for sustainable development. However, this diversity also presents a significant challenge in the form of linguistic barriers, given the importance of languages in communication.
The official language of Burkina Faso is French, and it has approximately 60 local languages [13].

This situation poses a challenge for communication and understanding among the different linguistic communities in the country, making it difficult for the majority of the population to access information. An effective solution to this situation is the automatic translation of local languages. Neural Machine Translation (NMT) is a rapidly evolving field fueled by advances in artificial intelligence (AI) and natural language processing (NLP). It is an architecture that allows machines to learn to translate between different languages [1]. However, Burkina Faso national languages have been underexplored in the field of neural machine translation, and the resources to do so are either non-existent or difficult to obtain, especially the data.
The overall objective of this study is to develop an efficient automatic translation system for Burkina Faso national languages, particularly "Mooré", to facilitate communication and understanding among speakers of these languages and other languages. To achieve this, we evaluated the effectiveness of various AI techniques for automatic translation of Burkina Faso national languages, collected and pre-processed a corpus of "Mooré" texts, as well as their translation into French.
Our work aims to promote linguistic inclusion in administrative, educational, and media spheres, starting with Moore. The rest of the article is organized as follows: in section 2, we provided a state of the art of works related to our objectives. In section 3, we presented our methodology. In section 4, we present our results and challenges. We conclude in section 5.

II RELATED WORK

Several recent works in the field of automatic language translation have been carried out, with the majority of them focusing on low-resource languages. In 2020, Dossou and Emezue [7] used an encoder-decoder architecture consisting of Gated Recurrent Units (GRU) to propose an automatic translation model for Fon, a language spoken in Benin, to French. They achieved a performance of 30.55 BLEU on the JW300 [5] and BeninLanguages datasets. The best results were obtained on data with diacritics (tonal marks). In the same year, Laura Martinus et al [8] proposed a translation model into English for six South African languages using the Transformer and achieved a score of 40 BLEU on the JW300 dataset. The authors demonstrated that the training data domain has an impact on model performance.

The Transformer is a neural network architecture based entirely on attention [3]. The authors of [3] introduced it in 2017 and showed that the Transformer outperforms encoder-decoder architectures based on recurrent neural networks for translation tasks on WMT2014 data. The absence of recurrent layers in the Transformer makes it faster to train.

In 2021, Hacheme [9] used the Transformer to propose a multilingual automatic translation model from English to Gbe (Fon and Ewe), known as English2GBE. The main goal was to demonstrate the benefits of a multilingual automatic translation model. They constructed three translation models: one for English to Ewe, one for English to Fon, and a multilingual model for English to Ewe and Fon (English2GBE). The results showed that the multilingual model outperformed the bilingual models. This was explained by the fact that the two languages are from the same family and share some characteristics, allowing them to learn from each other during model training.

The authors in [10] also demonstrated the effectiveness of the Transformer in translation tasks. They built two automatic translation models, one based on JoyNMT [6] and the other based on the Transformer. The Transformer-based models achieved better results. For their model training, they tested three data representation models and found that the Binary Pair Encoding (BPE) representation improved model performance. Tests were conducted using Bible data from YouVersion, JW300 data, and data provided by the South African government, Autshumato. The results were compared with the work of [8] and achieved a BLEU score at least 7 points higher.

Other researchers have used to pretrained models to train more powerful automatic translation models, despite the limited existing data. In 2019, the authors of [5] demonstrated that pretrained models have a significant impact on linguistic modeling, such as causal language modeling (CLM), masked language modeling (MLM), and translation language modeling (TLM). Pretraining multilingual language models leads to better results, especially in automatic translation tasks, where it achieved an average BLEU score of 75.1, compared to 71.5 for (Artetxe and Schwenk, 2018), which was the state of the art for the same language corpus translation task. Furthermore, they achieved a new state of the art with a BLEU score of 34.3 on WMT'16 German-English.

AfroLM [11] proposed a pretrained multilingual model on 23 African languages called AfroLM. This model is based on an active learning algorithm, which gives a model M the ability to query another model N to improve itself. In their case, they set M=N, making it a form of self-supervised active learning. They demonstrated that with 14 times less data, AfroLM is competitive with other pretrained language models, achieving an average F1-score of 80.13% with 0.73GB of data compared to 81.90% for AfroXLMR-Base with approximately 2.5 TB of data. Additionally, they proved that the model generalizes well for other NLP tasks. Their data was collected from news sources and covers various parts of the continent. They also used the BPE model for data representation.

III METHODOLOGY

In this section, we have described the methodology used in this document. This work is an extension of [12]. We begin by explaining how we collected and processed our data. Then, we describe how we have trained our language detection model, and finally, we present the results obtained in the next section. Figure 1 depicts our methodology.

3.1 Data collection and data processing

3.1.1 Data collection

We identified several data sources that contain text in "Mooré" with their translations in French. The first data source we used is the Jehovah's Witnesses' Bible from the jw.org[1] website, which provides translations of the Bible in "Mooré" and French. We collected both versions of the Bible directly from the website, treating each verse as a line in our dataset. We used web scraping for this task. Since we collected the entire Bible, we divided it into four (4) parts for each language, allowing us to run eight (8) tasks in parallel. This reduced the time required to approximately three (3) hours for both versions of the Bible, compared to six (6) hours for a single version without parallel processing.
Another data source is the ohchr website[2], which contains the Universal Declaration of Human Rights in both French and "Mooré". We also scraped this data, considering each sentence as a line in our dataset. Lastly, we utilized the "Mooré"-French dictionary index [3] in PDF format.

[1]https://www.jw.org/
[2]https://www.ohchr.org/fr
[3]https://www.webonary.org/moore/files/index-francais-moore.pdf

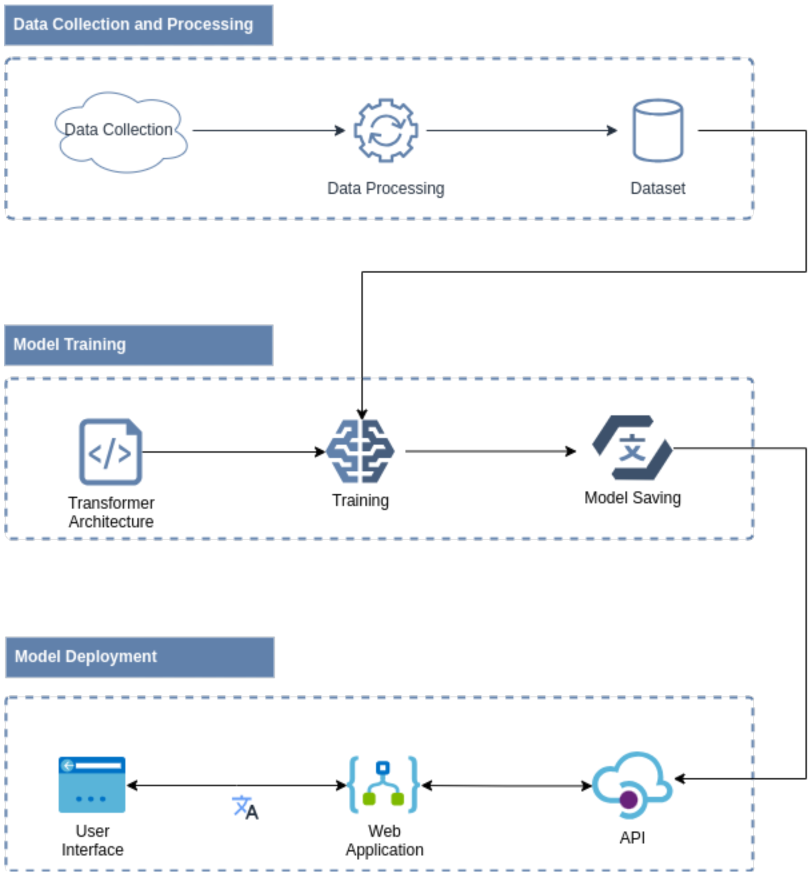

Figure 1: Comprehensive methodology for developing a Low-ressource language Translation system

We extracted data from this PDF document using data extraction techniques, resulting in a total of 36,178 lines for our dataset. Table I provides the number of lines obtained for each source.

Table 1: The number of lines for each data source

Source	JW	ohchr	index
Number of lines	31078	64	5036
Number of words (mos - fr)	820817 - 757509	2033 - 1527	5036 - 5036

3.1.2 Data processing

We cleaned the data to obtain quality data. We performed data alignment for the Universal Declaration of Human Rights with the assistance of three (3) individuals who aligned and then verified that others had aligned correctly on their side. For the JW data, we conducted verse-level alignment during data collection and observed that some values were expressed in numbers in one verse and in words in its translation. This inconsistency could lead to model comprehension issues. We identified these lines (1415 lines) and removed them from the dataset. For the index, we did not need to make any modifications to the initially collected version. Our final dataset comprises 34,763 lines ("Mooré" : mos, French : fr). We present a data excerpt in Table 2.

Table 2: Data excerpt

mos	fr
Maam a Poll sn yaa ned ning Kirist Zeezi sn b...	De la part de Paul , appelé pour être apôtre de...
n gls sebkãngã n tool Wnnaam tiging ning sn ...	à l'assemblée de Dieu qui est à Corinthe , à vo...
B bark la laaf sn yit Wnnaam sn yaa tõnd B...	Que Dieu notre Père et le Seigneur Jésus Christ...
Bala yãmb sn be a pg wã , yãmb paamda bũmb f...	En effet , par votre union avec lui vous avez é...

3.1.3 Tokenization

We used the SentencePiece tokenizer [4], a language-agnostic subword tokenizer. SentencePiece is a widely used tokenizer in Natural Language Processing (NLP) due to its linguistic versatility, ability to handle compound and rare words, flexibility, and strong performance. It works well with many languages, including "Mooré", offers customization options, is supported by various NLP frameworks, and is efficient for tokenization in various NLP tasks. We used the SentencePiece module implemented in TensorFlow[4] [5] with a vocabulary size (vocab_size) of 8,000 and set normalization to *false* to preserve accents, especially for "Mooré".

3.2 Model training

We divided our dataset as follows: 70% of the data for training, 20% for testing, and 10% for validation. We trained two machine translation models for "Mooré". The first model translates "Mooré" to French, and the second one translates French to "Mooré". We implemented the Transformer architecture described in [2] using TensorFlow. Each model was trained for 80 iterations on the training data, with an average of 13.5 minutes per iteration, totaling approximately 18 hours to train a model.

For the configuration of our models, we used four (4) layers ($num\_layers = 4$) and four (4) attention heads ($num\_heads = 4$). Our models have a dimension of 128 ($d\_model = 128$), hidden layer dimension of 512 ($dff = 512$). The dropout rate, which is the probability that a neuron is deactivated, is set to 10% ($dropout\_rate = 0.1$).The workspace we have on the server has the following characteristics:

- Processor: 4 CPUs with an average frequency of 3.0GhzRAM
- Memory: 64GB
- Operating system: Debian 5.10.140-1

IV RESULTS AND CHALLENGES

Table 3 shows the results we obtained after our initial training rounds. These results encourage us to continue our research work, even though we are severely limited in terms of linguistic resources, and the vast majority of our available data comes from biblical sources. The machine translation model from French to "Mooré" performs better than the model translating from "Mooré" to French. This indicates that the model is better at extracting context from French

[4] https://www.tensorflow.org/?hl=fr
[5] https://www.tensorflow.org/text/api_docs/python/text/SentencepieceTokenizer

sentences than from "Mooré" sentences.
"Mooré" is a tonal language, meaning that two words can have the same spelling but different meanings, as in "Saaga: balai ou pluie" (Saaga: broom or rain) or with a diacritical mark that also changes the meaning, as in "sãaga: diarrhée" (sãaga: diarrhea). These differences are more perceptible in spoken language than in written form, and this has an impact on the models' performance.

Table 3: Results of model evaluation with training, test and validation data

data	fr-mos
JW	72.61 BLEU
JW + index	71.38 BLEU
JW + index + ohdh	72.18 BLEU

We have developed a web application for machine translation to make it easier for the public to use. Figure 2 shows the translation interface from French to "Mooré".

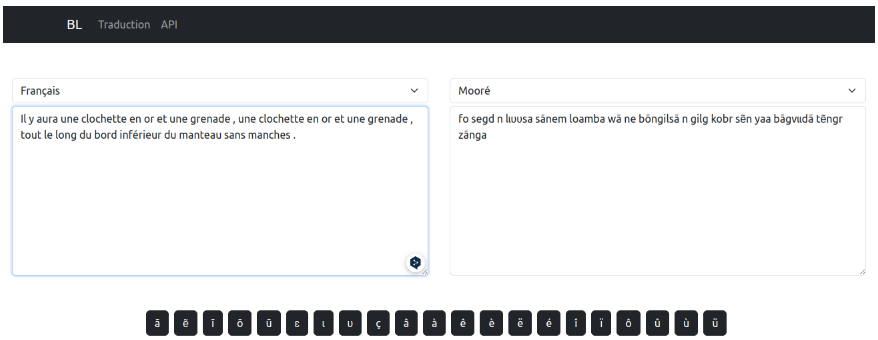

Figure 2: Translation from French to "Mooré"

Table 4 shows some examples of translations.

Table 4: Examples of French to "Mooré" Translations

setence	les troupes comptaient deux myriades de myriades de cavaliers — j'ai entendu leur nombre
correct	tãbbiisã sn yaa wedrdbã ra yaa tuspisi naoor tuspiiga mam wma b sõorã
predict	tãbbiisã sn yaa wedrdbã ra yaa tuspisi naoor tuspiiga mam wma b sõorã
setence	certains
correct	kere
predict	kere
setence	lenseignement technique et professionnel doit être généralisé
correct	tmminim la nus tm zãmsg kaorengã togame n piuugi
predict	tmminim la nus tm zãmsg kaorengã togame n piuugi

Table 5: Example of a mistranslation from French to "Mooré"

setence	quand joseph vit que son père gardait la main droite posée sur la tête d'éphraïm cela lui déplut il essaya donc de prendre la main de son père pour la déplacer de la tête d'éphraïm à la tête de manassé
correct	a zozf sn yã t'a ba wã kell n tika a efrayim zugã ne a nugrtgã pa ya noog ye d a makame n na n zk a ba wã nug a efrayim zug wã n t rogl a manase zug wã
predict	a zozf sn yã t ' a ba wã kell n tika a efrayim zugã ne a nugrtgã pa ya noog ye woto a makame n na n zk a ba wã nug a efrayim zug wã
setence	kapokier
correct	vaooka
predict	vaaga
setence	tout individu a droit à la vie à la liberté et à la sûreté de sa personne
correct	ned buud fãa tara sor n tõe n vnde n soog a menga la a gũnug a menga
predict	ned buud fãa tara sor n tõe n vnde n soog a menga la ned tõe n lebs n deega a vmã

V CONCLUSION

In this work, we trained two machine translation models based on the Transformer architecture. The first model facilitates automatic translation from 'Mooré' to French, while the second performs the inverse task, translating from French to 'Mooré'. Our models achieved respectable BLEU scores of 65.87 and 72.18, respectively. Ongoing research aims to refine these models by optimizing their parameters, and we have also developed a web application to make these machine translation capabilities accessible through a user-friendly interface.

As for future perspectives, we intend to compile a comprehensive dataset that pairs 'Mooré' with multiple other languages, with the goal of developing multilingual machine translation models. Additionally, we aim to address the unique challenges posed by the tonal nature of the 'Mooré' language, such as differentiating words with identical spellings but different meanings due to tonal variations. This will involve creating tonality-sensitive algorithms that can effectively interpret and generate accurate translations in the face of such complexities.

ACKNOWLEDGEMENT

This work was conducted as part of the Artificial Intelligence for Development in Africa (AI4D Africa) program, with the financial support of Canada's International Development Research Centre (IDRC) and the Swedish International Development Cooperation Agency (Sida).

References

[1] T. Luong, K. Cho, and C. D. Manning. "Neural Machine Translation". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. *Attention Is All You Need*. Dec. 5, 2017. arXiv: 1706.03762[cs].

[4] T. Kudo and J. Richardson. *SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing*. 2018. arXiv: 1808.06226 [cs.CL].

[5] Ž. Agić and I. Vulić. "JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pages 3204–3210.

[6] J. Kreutzer, J. Bastings, and S. Riezler. "Joey NMT: A Minimalist NMT Toolkit for Novices". 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): System Demonstrations*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pages 109–114.

[7] C. C. Emezue and F. P. B. Dossou. "FFR v1. 1: Fon-French neural machine translation". In: *Proceedings of the The Fourth Widening Natural Language Processing Workshop*. 2020, pages 83–87.

[8] L. Martinus, J. Webster, J. Moonsamy, M. S. Jnr, R. Moosa, and R. Fairon. "Neural machine translation for South Africa's official languages". In: *arXiv preprint arXiv:2005.06609* (2020).

[9] G. Hacheme. *English2Gbe: A multilingual machine translation model for {Fon/Ewe}Gbe*. Dec. 13, 2021. arXiv: 2112.11482[cs].

[10] T. J. Sefara, S. G. Zwane, N. Gama, H. Sibisi, P. N. Senoamadi, and V. Marivate. "Transformer-based Machine Translation for Low-resourced Languages embedded with Language Identification". In: *2021 Conference on Information Communications Technology and Society (ICTAS)*. 2021 Conference on Information Communications Technology and Society (ICTAS). Durban, South Africa: IEEE, Mar. 2021, pages 127–132. ISBN: 978-1-72818-081-6.

[11] B. F. P. Dossou, A. L. Tonja, O. Yousuf, S. Osei, A. Oppong, I. Shode, O. O. Awoyomi, and C. C. Emezue. *AfroLM: A Self-Active Learning-based Multilingual Pretrained Language Model for 23 African Languages*. 2022. arXiv: 2211.03263 [cs.CL].

[12] O. Hamed Joseph, S. Aminata, B. Delwende Eliane, K. Rodrique, K. Abdoul Kader, and B. Tégawendé F. "Neural Machine Translation for Mooré, a Low-Resource Language". In: ARIMA, May 2023.

[13] *Langues au Burkina Faso*. In: *Wikipédia*. Page Version ID: 201215951. Feb. 8, 2023.

[14] *Diversité culturelle et linguistique | African Declaration on Internet Rights and Freedoms*. URL: https://africaninternetrights.org/fr/principles/diversit%C3%A9-culturelle-et-linguistique (visited on 07/05/2023).