# Bayelemabaga: Creating Resources for Bambara NLP

**Allahsera Auguste Tapo**[1] **Kevin Assogba**[1]
**Christopher M. Homan**[1] **M. Mustafa Rafique**[1] **Marcos Zampieri**[2]
[1]Rochester Institute of Technology [2]George Mason University
{aat3261, kta7930, cmhvcs, mmrvcs}@rit.edu mzampier@gmu.edu

## Abstract

Data curation for under-resource languages enables the development of more accurate and culturally sensitive natural language processing models. However, the scarcity of well-structured multilingual datasets remains a challenge for advancing machine translation in these languages, especially for African languages. This paper focuses on creating high-quality parallel corpora that capture linguistic diversity to address this gap. We introduce Bayelemabaga, the most extensive curated multilingual dataset for machine translation in the Bambara language, the vehicular language of Mali. The dataset consists of 47K Bambara-French parallel sentences curated from 231 data sources, including short stories, formal documents, and religious literature, combining modern, historical, and indigenous languages. We present our data curation process and analyze its impact on neural machine translation by fine-tuning seven commonly used transformer-based language models, i.e., MBART, MT5, M2M-100, NLLB-200, Mistral-7B, Open-Llama-7B, and Meta-Llama3-8B on Bayelemabaga. Our evaluation on four Bambara-French language pair datasets (three existing datasets and the test set of Bayelemabaga) show up to $+4.5$, $+11.4$, and $+0.27$ in gains, respectively, on BLEU, CHRF++, and AfriCOMET evaluation metrics. We also conducted machine and human evaluations of translations from studied models to compare the machine translation quality of encoder-decoder and decoder-only models. Our results indicate that encoder-decoder models remain the best, highlighting the importance of additional datasets to train decoder-only models.

## 1 Introduction

Driven by the availability of massive, digitized data sets and advancements in neural architectures (Chernyavskiy et al., 2021), state-of-the-art natural language processing (NLP) models are widely applied to the world's high-resource languages (e.g., English, French, Spanish). They are employed in tasks such as machine translation (MT) (Wu et al., 2016), name entity recognition (NER) (Jehangir et al., 2023), automatic speech recognition (ASR) (Radford et al., 2023).

Yet the vast majority of the world's languages, and by extension the people who speak these languages, lack the digitized data resources needed to support MT systems (Weeks, 2021) such as Google Translate and other important language technology applications. These under-resourced languages have yet to benefit from recent advances because they lack the large volumes of text needed to drive language technology development. The case of neural machine translation (Gu et al., 2018) is particularly representative as it requires large volumes of parallel data between pairs of source and target languages. Moreover, the available data in under-resource languages is often noisy and diverse, with non-standardized spelling, accenting, marking, multiple scripts, code-switching, etc.

For example, Bambara, a tonal language with a rich morphology from the Mande language family[1], has several competing writing systems: Adjami (Arabic-based), Latin, and N'ko. However, as a historically oral-only language, most Bambara speakers have never been taught to read or write it. As a rule, available resources don't fit standard writing systems (e.g., systems developed during colonization) or lack standard orthography or ways to express features, such as tonality, absent in colonial scripts.

To cope with this problem, language experts are actively working on standardizing the existing vocabulary and coining new words to enrich the language and support automated text processing. Initiatives in this direction include Masakhane for

[1]https://en.wikipedia.org/wiki/Mande_languages



*Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12060–12070

African languages (Orife et al., 2020), the increased presence of under-resourced languages in the popular machine translation (MT) competitions of the annual conference on machine translation (WMT) (Barrault et al., 2019, 2020), and AfricaNLP, a workshop dedicated to African language technologies.

To help alleviate the scarcity of data for machine translation of under-resourced languages, we introduce Bayelemabaga, a new comprehensive dataset for machine translation that comprises 46,976 pairs of Bambara and French sentences. We collected data from decades of linguistics work on Bambara from INALCO[2]'s Corpus Bambara de Reference[3], aligned collected sentences in both languages, investigated their morphological structure, and curated the content to ensure adequacy for machine translation.

We evaluate the adequacy of Bayelemabaga by answering the following research questions (RQ):

- RQ1: How does the quality of our curated dataset compare to the raw data?
- RQ2: What is the impact of our dataset on improving translation results compared to existing models fine-tuned on the currently scarce data?
- RQ3: How do emerging large language models, which were not fine-tuned for machine translation for the Bambara language, perform when evaluated after fine-tuning with our dataset?

Bayelemabaga aims to improve the quality of translation models for the Bambara language by providing a richer training resource and adaptability to other natural language processing tasks.

## 2 Related Work

Several linguistic studies have been conducted on the Bambara language, providing valuable insights into its structure (Ermisch, 2013), syntax (Bird, 1966), grammar (Dombrowsky-Hahn, 2020), and phonology (Green, 2010). These studies serve as foundational resources for further research and resource development (Vydrin, 2009; Vydrin et al., 2011; Vydrin, 2013, 2014; Vydrine, 2015; Vydrin et al., 2016; Vydrin, 2018). While these linguistic studies are essential for understanding language, more up-to-date and accessible resources that can be utilized by a broader audience, including language learners, educators, researchers from the NLP community, and the general public, are needed.

Educational materials for learning Bambara are relatively scarce compared to more widely taught high-resource languages, such as French or English. However, some resources exist, primarily in textbooks and language learning guides (Bird and Kante, 1976). While these materials are valuable, they may be outdated or difficult to access, particularly for learners outside academic or linguistics research settings. There is a need for more interactive and accessible educational resources that cater to different learning styles and proficiency levels.

Some online dictionaries and language learning apps exist (Vydrin, 2013) but are often limited in scope or functionality. Additionally, there is a lack of digital corpora or databases that could facilitate machine translation (MT), automatic speech recognition (ASR), and text-to-speech (TTS) (Tapo et al., 2020). Leveraging technology to create digital resources, such as interactive language learning platforms, mobile apps, and multimedia content, could significantly improve accessibility and engagement for Bambara learners and speakers.

Various organizations and initiatives have been working to promote and preserve the Bambara language, such as INALCO and the Academie Malienne des Langues (AMALAN) in Mali, which aims to standardize and encourage the use of national languages, including Bambara. However, there is a need for more comprehensive and sustained efforts to create resources that support language preservation, such as the development of educational materials, the promotion of Bambara in media and literature, and the integration of the language into formal education systems (Daou et al., 2024; Daou and Mohanty, 2024). Additionally, while the Bambara language has a rich linguistic heritage and a significant number of speakers, the availability of resources for neural machine translation is limited compared to high-resource languages like French or English (Akhbardeh et al., 2021).

To address the gaps and meet the growing demand to enable Bambara to be a human-technology language, we put together a collaborative effort involving linguists, educators, technology experts, and community stakeholders to curate decades of linguistic data from varying sources, including books, periodicals, news, etc., for machine learning, including machine translation.

[2] http://www.inalco.fr/en
[3] http://cormand.huma-num.fr/

## 3 The Bayelemabaga Dataset

We created a parallel text dataset for the dialect continuum of Mande languages spoken in West Africa. Our contribution focuses on the Bambara language, described by Tapo et al. (2020) as a tonal language (different words with different inflections convey different meanings) with a rich morphological structure, similar to other languages in the Mande language family. This family consists of several languages (Bambara, Dyula, Maninka, etc.) spoken by 30–40 million people across the African continent, among whom there are around 15-18 million Bambara speakers, primarily in Mali. With three central writing systems (Adjami, Latin, and N'ko), Bambara uses diacritical marks to indicate high or low tones in the spoken language, helping distinguish between words that use the same sequence of letters. There are 27 letters in its Latin writing script except for $q, v$, and $x$, commonly seen in French and English. An example of an additional character is ɛ, as shown in the Bambara translation of the following phrase.

**English:**
the useful homemade medication

**Bambara:**
farafinfura minnu bɛ se ka bana furakɛ

**French:**
les remèdes maison utiles

To ensure a standard orthography across our dataset, we invited a dozen Bambara linguists and language experts for a workshop to establish a unified scientific orthographic system for the Bambara language. The workshop addressed the challenges of Bambara's orthographic variation, which stemmed from its multiple dialects and writing systems. The resulting orthographic system provides a foundation for Bambara text processing by identifying clear syntax rules to ensure the effectiveness of our dataset for natural language understanding.

We introduce Bayelemabaga, a collection of 46,976 Bambara-French parallel sentences compiled from various sources, including the Dokoto Project (The Dokotoro Project), religious books, SIL dictionary sentences, and the Corpus Bambara de Référence (Vydrin et al., 2011). The dataset was collected for decades-long linguistics work and curated following the rules developed during the workshop mentioned above. After data collection, we ensured all sentences were written in Latin script before proceeding to data curation and sentence alignment. Bayelemabaga is ready-to-use for MT tasks and available on open-source platforms [4].

### 3.1 Data Collection

The Bayelemabaga dataset was curated from 231 data sources, ranging from informal content (short stories, blog posts) to formal documents with modern language structures (books, news releases) to religious literature (passages from the Bible and the Quran). The initial data collection process was manual and led to the accumulation of datasets with four distinct classes reflecting different levels of annotation described as follows:

- *Non-Annotated Bambara with French Translation*: This subset contains raw Bambara sentences without linguistic annotations, paired with their corresponding French translations. These translations were directly sourced from available resources and have not been modified or adjusted by linguistic experts.

- *Annotated Bambara with French Translation*: This subset contains linguistically annotated Bambara sentences paired with their French translations. The annotations in Bambara include syntactic and morphological features of the language intended to provide deeper linguistic insight. The French translations in this subset were also sourced from available resources and may not fully align with the nuances captured by the Bambara annotations.

- *Annotated Bambara with Two French Translations (Original and Adjusted)*: In this subset, annotated Bambara sentences are provided along with two French translations: the original translation (as sourced from the web) and an adjusted translation, revised by linguists experts to reflect the nuances of the Bambara annotations. The adjusted translations were designed to correct any inaccuracies or ambiguities in the original translation, ensuring closer fidelity to the Bambara source.

- *Annotated Bambara with Adjusted French Translation*: The final subset contains annotated Bambara sentences with only the

[4] https://robotsmali-ai.github.io/datasets/

adjusted French translations. The original French translations were not included in this subset, as the adjusted translations provide a higher-quality parallel corpus for MT tasks.

### 3.2 Data Alignment and Curation

Working with our linguistic and language experts, we combined data from all four classes described in Section 3.1 and employed a three-step data alignment and curation process.

First, the linguists and language experts divided the Bambara and French sentences into separate files with one sentence per line. In a third file, they maintained a mapping of lines in the Bambara file with the equivalent translation in the French file. This mapping is represented as "$nTABm$", where $n$ is the Bambara sentence line number, $m$ is the French sentence line number, and "$TAB$" delimits $n$ from $m$. We indicated lines where the sentence does not have a corresponding translation by "-1" standing for "no matching."

Next, we performed automatic alignment of the sentences by parsing the mappings file to generate a JSON file that combined each French sentence with its equivalent Bambara translation into one dictionary object. To ensure that every sentence has a reference translation, we skipped all instances with "$nTAB-1$" and "$-1TABm$" mappings.

Finally, we used Python's regular expression (re) library to clean both the Bambara and French files. This removed unwanted elements such as HTML/XML tags, non-printable characters, and orphan symbols. We also eliminated newline characters within aligned pairs of French and Bambara sentences.

### 3.3 Data Partitioning

To ensure models trained with our dataset are robust and generalize well to new data, Bayelemabaga was divided into training, validation, and testing sets with a ratio of 80%, 10%, and 10% of all curated sentences, respectively. As a result, the number of parallel sentences in the training, validation, and test sets is $37,580$, $4,698$, and $4,698$, respectively. Table 1 details each of the three sets, including the number of sentences, the average sentence length, the total number of tokens, and unique tokens.

Table 1: Overview of the Bayelemabaga dataset including Number of sentences (N. Sen.), Average sentence length (Avg. SLen.), Number of tokens (N. Tok.) and Number of unique tokens (N. UTok.)

| Partition | N. Sen. | Avg. SLen. | N. Tok. | N. UTok. |
|---|---|---|---|---|
| Training | 37,580 | 15 | 531,501 | 37,975 |
| Validation | 4,698 | 16 | 73,727 | 8,388 |
| Testing | 4,698 | 10 | 44,278 | 5,516 |

## 4 Experiments

We evaluate the quality of the Bayelemabaga dataset by comparing its performance before and after curation using various machine translation models. We also investigate the contributions of our curated dataset in improving the state-of-the-art performance of machine translation in the Bambara language by combining our newly collected dataset with existing ones and examining the overall performance of selected models.

### 4.1 Evaluation Setup

Our experimental testbed comprises computing resources from Rochester Institute of Technology's Research Computing facility (Rochester Institute of Technology, 2024). The computing cluster has 64 nodes with two 2.7 GHz Intel Xeon Gold 6150 processors (36 cores), 384 GB of RAM, two 100 Gb/s Ethernet network connections, and 7 TB external storage exposed through a parallel file system. Our experiments were executed on a single node with four NVIDIA A100 GPUs (40 GB high-bandwidth memory each). However, each model was fine-tuned and evaluated on a dedicated GPU.

### 4.2 Methods

We evaluate two transformer-based language model architectures (encoder-decoder and decoder-only), instrumenting seven models for machine translation to generate text in Bambara or French, depending on the evaluation source language.

#### 4.2.1 Evaluation Data

Our analyses focus on four different datasets, described in Table 2: (i) *Dictionary* consists of a set of dictionary entries, each of a single sentence, in Bambara and translated into French and English (Tapo et al., 2020). (ii) *Medical* is a collection of health guidance in French, English, and Bambara (Tapo et al., 2020). (iii) *News* is a set of translations of news from French into Bambara (Adelani et al., 2022). (iv) *Bayelemabaga* is our curated and aligned dataset (§ 3). We also use the version of the dataset before curation to assess the quality of our curation and alignment effort (§ 4.3.1).

Table 2: Overview of the datasets. Dictionary, Medical, News, and Bayelemabaga datasets and their splits.

| Dataset | Train | Dev | Test |
|---|---|---|---|
| Dictionary | 1,521 | 265 | 266 |
| Medical | 2,973 | 454 | 456 |
| News | 3,013 | 1,500 | 1,500 |
| Bayelemabaga | 37,580 | 4,698 | 4,698 |

Table 3: Finetuning Hyperparameters.

| **Parameter** | **Value** | **Parameter** | **Value** |
|---|---|---|---|
| Learning Rate | $2e^{-4}$ | Max Seq. Length | 80 |
| Weight Decay | $1e^{-3}$ | Max Grad. Norm. | 0.3 |
| Max Epochs | 3 | QLoRA Attention | 64 |
| Warmup Ratio | 0.03 | QLoRA Alpha | 16 |
| Optimizer | Adam8bit | QLoRA Dropout | 0.1 |

#### 4.2.2 Models

**Encoder-Decoder** We experiment with three encoder-decoder models fine-tuned with datasets from 16 African languages by Adelani et al. (Adelani et al., 2022), as well as a version of the NLLB-200 model previously fine-tuned for French-Bambara MT: (1) *MBART*. A fine-tuned version of the MBART model on African languages, tailored for sequence-to-sequence multilingual tasks with strong translation and text generation capabilities across various languages. (2) *MT5*. a multilingual variant of the T5 model employed for various tasks, including translation, summarization, and question-answering tasks. (3) *M2M-100*. a multilingual MT model designed to handle many-to–many language translations between any pair of 100 languages. (4) *NLLB-200* [5]. A fine-tuned version of the No Language Left Behind (NLLB-200) distilled 600M variant for French-Bambara MT.

**Decoder-only** We further explore the quality of Bayelemabaga for MT tasks using emerging transformer-based language model architectures that only feature decoders. Although these models are designed for general-purpose language generation tasks, several existing efforts have adapted them for specific tasks, including machine translation (Adelani et al., 2024; Tonja et al., 2024). We selected the following three open-source models: (1) *Open-Llama-7B*. an open-source adaptation of the LLaMA model, trained for general-purpose language understanding and generation. (2) *Mistral-7B*. a language model known for high performance and efficiency in text generation and comprehension. We used a variant with an extended vocabulary of 32 KB. (3) *Meta-Llama3-8B*. a language model developed by Meta AI and optimized for various NLP tasks.

[5] https://huggingface.co/oza75/nllb-600M-mt-french-bambara

#### 4.2.3 Hyperparameters

We present the hyperparameters used to fine-tune the selected models in Table 3. Most hyperparameters for encoder-decoder models follow existing work on machine translation for the Bambara language (Adelani et al., 2022). Default hyperparameters from Huggingface Transformers were adopted if not included in Table 3.

#### 4.2.4 Evaluation Metrics

We compare the quality of different MT systems using widely-known n-gram matching evaluation metrics, SacreBLEU (BLEU) (Post, 2018) and CHRF++ (Popović, 2015). We also use AfriCOMET (Pu et al., 2021), a learned COMET metric for MT covering 13 African languages. We validate our results by reporting the true mean score estimated from bootstrap resampling and the 95% confidence interval around the mean using a bootstrap resample size of 1000 samples.

### 4.3 Performance Results

#### 4.3.1 Automatic Evaluation

**Performance Benefits of Bayelemabaga:** Our first set of experiments investigates the contributions of our newly curated dataset, Bayelemabaga, to machine translation for under-resourced languages, especially the Bambara language. We achieve this objective by comparing the scores of three evaluation scenarios, each characterized by the state of the models. We consider the models introduced in Section 4.2.2 as baselines and compared against two fine-tuned versions, respectively, on the raw version of our collected dataset and our curated dataset, i.e., Bayelemabaga. The scores are reported in Table 4. We found that fine-tuning with the Bayelemabaga dataset enhances the quality of generated translations, especially for MBART, MT5, and NLLB-200 models. The score of the pre-trained M2M-100 model is slightly higher on the Dictionary, Medical, and News datasets. However, we can conclude that pre-trained and fine-tuned versions of M2M-100 have comparable quality based on their standard deviations.

**Alignment and Curation Quality Assessment:** Next, we evaluate the quality of our data alignment

Table 4: Evaluation scores of pre-trained and fine-tuned encoder-decoder models on out-of-domain and in-domain datasets. Models were either fine-tuned on the training set of the raw dataset or Bayelemabaga and evaluated on test sets from three out-of-domain datasets (Dictionary, Medical, and News) and Bayelemabaga.

| | Dictionary | Medical | News | Bayelemabaga |
|---|---|---|---|---|
| | | BLEU | | |
| MBART (Pre-trained) | 0.1 ± 0.1 | 1.3 ± 1.3 | 1.3 ± 1.1 | 0.3 ± 0.1 |
| MBART (Raw) | 4.8 ± 3.0 | 4.3 ± 2.2 | **11.3 ± 2.4** | 1.2 ± 0.2 |
| MBART (Bayelemabaga) | **17.4 ± 12.4** | **10.1 ± 5.2** | 7.2 ± 3.1 | **2.6 ± 1.1** |
| MT5 (Pre-trained) | 0.4 ± 0.3 | 0.9 ± 1.1 | 2.4 ± 1.3 | 0.5 ± 0.2 |
| MT5 (Raw) | 0.2 ± 0.1 | 1.0 ± 1.0 | 1.9 ± 1.2 | 0.3 ± 0.1 |
| MT5 (Bayelemabaga) | **3.4 ± 1.6** | **3.9 ± 2.0** | **2.7 ± 1.1** | **1.6 ± 0.7** |
| M2M-100 (Pre-trained) | **27.8 ± 7.4** | **7.6 ± 2.2** | **18.0 ± 3.1** | 4.6 ± 1.6 |
| M2M-100 (Raw) | 26.0 ± 7.5 | 6.9 ± 2.4 | 12.5 ± 2.9 | 5.0 ± 1.9 |
| M2M-100 (Bayelemabaga) | 25.0 ± 6.3 | 7.2 ± 1.9 | 11.6 ± 2.0 | **6.4 ± 2.0** |
| NLLB-200 (Pre-trained) | 29.3 ± 7.5 | 10.1 ± 2.4 | 14.1 ± 2.5 | **13.2 ± 3.7** |
| NLLB-200 (Raw) | 27.4 ± 9.8 | 7.7 ± 2.2 | 14.2 ± 3.0 | 8.0 ± 3.0 |
| NLLB-200 (Bayelemabaga) | **32.1 ± 7.4** | **10.6 ± 2.5** | **14.6 ± 2.5** | 12.5 ± 3.9 |
| | | CHRF++ | | |
| MBART (Pre-trained) | 2.0 ± 0.6 | 6.1 ± 1.3 | 10.8 ± 2.0 | 5.4 ± 1.6 |
| MBART (Raw) | 23.8 ± 3.2 | 15.9 ± 1.3 | 34.6 ± 2.6 | 17.8 ± 2.1 |
| MBART (Bayelemabaga) | **42.1 ± 9.6** | **25.6 ± 5.0** | **36.5 ± 2.9** | **23.3 ± 3.9** |
| MT5 (Pre-trained) | 7.8 ± 2.0 | 4.7 ± 0.6 | 12.5 ± 2.0 | 7.5 ± 1.0 |
| MT5 (Raw) | 5.3 ± 1.1 | 3.7 ± 0.3 | 9.0 ± 1.7 | 5.2 ± 0.8 |
| MT5 (Bayelemabaga) | **17.0 ± 3.2** | **11.0 ± 1.6** | **13.9 ± 1.7** | **16.6 ± 1.8** |
| M2M-100 (Pre-trained) | 48.2 ± 5.7 | 24.3 ± 1.9 | **43.6 ± 2.7** | 24.3 ± 1.8 |
| M2M-100 (Raw) | 45.3 ± 6.4 | 21.2 ± 2.0 | 33.4 ± 3.7 | 23.6 ± 1.8 |
| M2M-100 (Bayelemabaga) | **47.7 ± 5.2** | **24.4 ± 1.7** | 35.7 ± 2.3 | **25.8 ± 2.0** |
| NLLB-200 (Pre-trained) | 47.9 ± 5.6 | 32.3 ± 2.2 | 38.3 ± 2.0 | **33.2 ± 3.1** |
| NLLB-200 (Raw) | 45.5 ± 6.5 | 30.8 ± 2.2 | 37.3 ± 2.6 | 29.4 ± 2.8 |
| NLLB-200 (Bayelemabaga) | **50.8 ± 5.6** | **33.2 ± 2.1** | **38.4 ± 2.1** | 33.0 ± 3.2 |
| | | AFRICOMET | | |
| MBART (Pre-trained) | 0.32 | 0.22 | 0.27 | 0.30 |
| MBART (Raw) | 0.53 | 0.36 | 0.50 | 0.51 |
| MBART (Bayelemabaga) | **0.65** | **0.45** | **0.53** | **0.57** |
| MT5 (Pre-trained) | 0.22 | 0.09 | 0.23 | 0.21 |
| MT5 (Raw) | 0.11 | 0.07 | 0.12 | 0.10 |
| MT5 (Bayelemabaga) | **0.37** | **0.22** | **0.30** | **0.37** |
| M2M-100 (Pre-trained) | 0.66 | 0.47 | **0.57** | **0.55** |
| M2M-100 (Raw) | 0.63 | 0.41 | 0.49 | 0.53 |
| M2M-100 (Bayelemabaga) | 0.66 | 0.47 | 0.54 | 0.53 |
| NLLB-200 (Pre-trained) | 0.66 | 0.50 | 0.56 | 0.53 |
| NLLB-200 (Raw) | 0.63 | 0.47 | 0.55 | 0.52 |
| NLLB-200 (Bayelemabaga) | 0.66 | **0.53** | 0.56 | 0.53 |

and curation process. Table 5 reports the comparison results between models fine-tuned on the collected data (Raw) and Bayelemabaga. These results show that Bambara-aware alignment and curation improve translation quality across all models evaluated on the test set of our curated data. We observe up to $+4.5$, $+11.4$, and $+0.27$ in gains, respectively, on BLEU, CHRF++, and AfriCOMET scores. Meanwhile, the fine-tuning of MT models on the raw data shows better scores when evaluated on the raw data. These observations are as expected because fine-tuning the models on the raw dataset forced them to learn from numerous sentences with uninformative and meaningless tokens, making the fine-tuned models less likely to generate logical and structured translations. Similarly, models fine-tuned on the curated data could not reproduce the low-quality translation expected on the raw dataset.

**Bi-directional Machine Translation:** Table 6 reports the evaluation scores of encoder-decoder and decoder-only models for both $fr \rightarrow bam$ and

Table 5: Evaluation scores of encoder-decoder models fine-tuned on raw or curated datasets. Training sets were used during fine-tuning, and test sets were used during evaluation in each category.

| Models | Raw Dataset | | | Curated Dataset (Bayelemabaga) | | |
|---|---|---|---|---|---|---|
| | BLEU | CHRF++ | AFRICOMET | BLEU | CHRF++ | AFRICOMET |
| MBART (Raw) | **7.1 ± 2.5** | 13.5 ± 1.9 | 0.15 | 1.2 ± 0.2 | 17.8 ± 2.1 | 0.51 |
| MBART (Curated) | 2.2 ± 1.7 | **14.9 ± 3.6** | **0.23** | **2.6 ± 1.1** | **23.3 ± 3.9** | **0.57** |
| MT5 (Raw) | **13.3 ± 5.5** | **11.5 ± 1.6** | **0.11** | 0.3 ± 0.1 | 5.2 ± 0.8 | 0.10 |
| MT5 (Curated) | 2.0 ± 1.4 | 6.4 ± 2.9 | 0.05 | **1.6 ± 0.7** | **16.6 ± 1.8** | **0.37** |
| M2M-100 (Raw) | **8.1 ± 2.2** | **15.6 ± 2.2** | 0.25 | 5.0 ± 1.9 | 23.6 ± 1.8 | 0.53 |
| M2M-100 (Curated) | 1.0 ± 0.7 | 16.4 ± 3.9 | **0.26** | **6.4 ± 2.0** | **25.8 ± 2.0** | 0.53 |
| NLLB-200 (Raw) | **4.7 ± 1.1** | 14.3 ± 1.5 | 0.21 | 8.0 ± 3.0 | 29.4 ± 2.8 | 0.52 |
| NLLB-200 (Curated) | 2.4 ± 1.4 | **17.5 ± 4.6** | **0.28** | **12.5 ± 3.9** | **33.0 ± 3.2** | **0.53** |

Table 6: Evaluation scores of encoder-decoder and decoder-only LLMs fine-tuned on the training set of all datasets (Dictionary, Medical, News, and Bayelemabaga) and evaluated on the test set in both source-target directions. Reported results for decoder-only models are of zero-shot inference.

| Models | $fr \rightarrow bam$ | | | $bam \rightarrow fr$ | | |
|---|---|---|---|---|---|---|
| | BLEU | CHRF++ | AFRICOMET | BLEU | CHRF++ | AFRICOMET |
| **Encoder-Decoder** | | | | | | |
| MBART | 30.2 ± 8.0 | 50.5 ± 5.9 | 0.66 | 21.8 ± 6.9 | 38.3 ± 5.3 | 0.42 |
| MT5 | 5.8 ± 2.6 | 21.8 ± 3.5 | 0.44 | 9.0 ± 3.8 | 24.5 ± 3.8 | 0.30 |
| M2M-100 | 33.1 ± 7.7 | 51.8 ± 5.6 | 0.66 | 32.7 ± 6.8 | 49.3 ± 5.6 | 0.52 |
| NLLB-200 | **34.2 ± 7.4** | **54.0 ± 5.6** | **0.69** | **34.1 ± 7.5** | **51.4 ± 5.6** | 0.56 |
| **Decoder-only** | | | | | | |
| Mistral-7B | 1.1 ± 0.5 | 15.0 ± 1.2 | 0.51 | 1.0 ± 1.1 | 8.7 ± 0.9 | **0.61** |
| Open-Llama-7B | 1.0 ± 0.5 | 11.7 ± 1.1 | 0.46 | 0.3 ± 0.2 | 7.5 ± 0.8 | 0.55 |
| Meta-Llama3-8B | 1.5 ± 0.7 | 12.4 ± 1.1 | 0.48 | 0.1 ± 0.1 | 8.4 ± 0.9 | 0.41 |

$bam \rightarrow fr$ translation. Our results show that encoder-decoder models outperform decoder-only models. The performance gap could be due to encoder-decoder architectures' distinct encoding and decoding phases, which better equip the models to revisit knowledge acquired from the encoding phase to handle the linguistic complexities of the translation task. However, decoder-only models maintain AfriCOMET scores that are competitive with the encoder-decoder models, showing that both models generate translation with a comparable correlation with human judgment, especially in under-resourced African languages.

**Impact of zero-shot and few-shot translations on Decoder Models:** We also investigate the performance improvement offered to decoder models by $k$-shot learing for $k \in \{0, 1, 3, 5\}$. See Figure 1. Compared to CHRF++, the BLEU metric shows high variability across different shots, suggesting inconsistent translation quality. CHRF++ better captures the morphological features of Bambara, while AfriCOMET offers a stronger correlation to human judgment. Based on the latter two metrics, all three decoder models achieve the highest evaluation scores with zero-shot prompting. The zero-shot performance can be attributed to a strong generalization from pre-training, as all three models were pre-trained on massive multilingual datasets, including languages that expose patterns similar to Bambara. Combined with our carefully curated dataset, the models learned a more comprehensive representation of the Bambara language, achieving comparable scores without additional examples. While the AfriCOMET scores are competitive with encoder-decoder models, the difference between BLEU and CHRF++ scores requires further investigations into how evaluated models represent language from the perspective of native Bambara speakers.

### 4.3.2 Human Evaluation

We conducted a human evaluation using a sample of 100 translations from each of the seven models. Our evaluation approach consisted of randomly selecting samples from all translated sentences and matching the same selected samples for all models. We defined human evaluation metrics to assess the quality and underlying problems in the model translation. We evaluated the quality of the translation

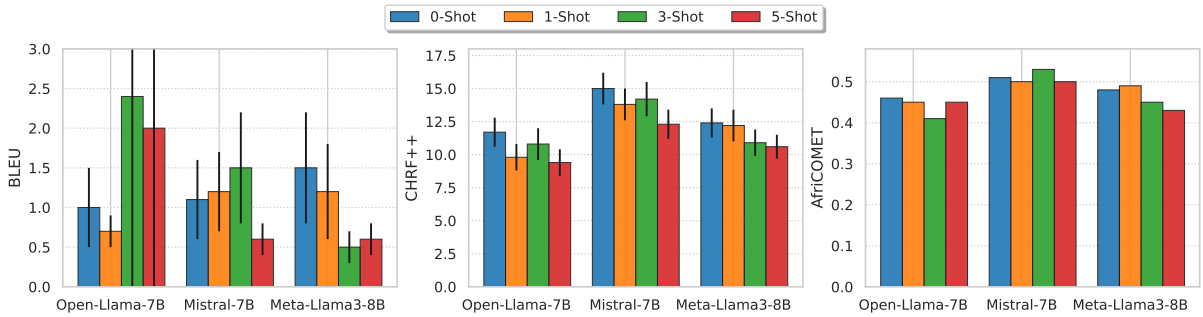


Figure 1: Evaluation of decoder models on translations from French to Bambara with 0-shot, 1-shot, and few-shots.

Table 7: Summary of human evaluation from native Bambara speakers on translations from by evaluated models.

| Models | Avg. (Score 1-5) | Quality (good/bad) | Problem (adequacy/fluency/understanding) |
|---|---|---|---|
| MBART | 3.54 | 65/35 | 25/5/15 |
| MT5 | 3.41 | 61/39 | 27/10/13 |
| NLLB-200 | **4.49** | **93/7** | **15/5/3** |
| M2M-100 | 3.9 | 73/27 | 24/6/9 |
| MISTRAL | 1.02 | 0/100 | 0/0/100 |
| Open-Llama-7B | 1 | 0/100 | 1/0/99 |
| META-Llama3-8B | 1.06 | 1/99 | 2/0/97 |

Table 8: Details insights on example translations and corresponding human evaluation scores.

| No. | Source | Reference | Model | Output | Quality | Problem | Score |
|---|---|---|---|---|---|---|---|
| 1 | Il a de la chance | A kunna ka di. | Open-Llama-7B | A bolo, a bolo. | Bad | Understanding | 1 |
| 2 | Il a de la chance | A kunna ka di. | MT5 | Jigi b'a la. | Good | Adequacy | 4 |
| 3 | eau froide | ji suma | M2M-100 | ji sumalen | Good | | 5 |

as *good* or *bad* based on problems related to their *fluency*, *adequacy*, and *understanding*. The *fluency* metric evaluates how easily readable the translation is; *adequacy* focuses on whether the translation contains ideal word choices to convey the desired meaning; and *understanding* evaluates if the translation makes sense to the Bambara speaker. We then introduced a score of one to five to rank all models' overall performance, allocating a score of five to the model that depicts the best quality with no problem. To conduct the human evaluation, we recruited two human evaluators and presented them with the translations without providing information on the corresponding model.

Table 7 summarizes our human evaluation results, reporting on the model score, the number of samples of good or bad quality, and the number of samples with problems related to fluency, adequacy, or understanding. We found that the NLLB-200 model performs best, with an average score of 4.58. All evaluated translations were good, but one sample had a fluency problem, and three had an adequacy problem.

Table 8 highlights three evaluated samples and their corresponding human evaluation. Example 1 generated by Open-Llama-7B contains Bambara words that are not correct translations for the source sentence "Il a de la chance", which translates into English as "He is lucky". The model translates as "A bolo" (meaning "in its hand" in English), which does not match the correct translation. Another example is 3, a translation of "eau froide" ("cold water" in English) into "ji sumalen" by the M2M-100 model. The additional phrase "len" does not appear in the reference but is used in the Bambara community, validating the machine translation.

The human evaluation scores support the machine translation results in Table 6, where the best models are encoder-decoder architectures, with NLLB-200 generating the best translations. Translations from decoder-only models scored worst as they could not rank as good. The models generate sentences that may have a meaning in the target language but do not correctly translate the source text. In multiple scenarios, they fail to combine words of the target language into a meaningful sentence, which explains their lower scores than encoder-decoder models.

## 5 Conclusion

In this paper, we introduced Bayelemabaga, a Bambara-French parallel corpora of 47K pairs of sentences collected from 231 data sources and curated to improve the quality of MT tasks. We explored the effect of curated data on MT compared to utilizing the raw dataset. We observed that Bayelemabaga improves translation quality by up to $+4.5$, $+11.4$, and $+0.27$ on BLEU, CHRF++, and AfriCOMET scores. Furthermore, we investigated the benefits of introducing a new data set by fine-tuning seven MT models (MBART, MT5, M2M-100, NLLB-200, Mistral-7B, Open-Llama-7B, and Meta-Llama3-8B) on Bayelemabaga, and evaluating them on three existing Bambara-French corpora. Our comparisons demonstrated that models fine-tuned on Bayelemabaga improve the translation quality across all datasets. We also explored the impact of our new dataset on existing encoder-decoder and decoder-only models. Machine and human evaluations showed that the encoder-decoder models yield the highest quality in Bambara-French and French-Bambar translations. In future work, we will conduct a more detailed human evaluation to explain the performance of decoder-only models and investigate if our observations on machine translation apply to speech data, especially for primarily oral languages (POLs).

## Acknowledgements

This work was carried out with support from Lacuna Fund, an initiative co-founded by The Rockefeller Foundation. It was also partly supported by the Google PhD Fellowship awarded to the first author, Allahsera Auguste Tapo, and Canada's International Development Research Centre We would also like to thank the staff at INALCO and Robots-Mali.

## Limitations

Two significant issues for machine translation are ambiguity and non-standard speech (Berthouzoz, 1999; Koehn and Knowles, 2017). This work does not directly address disambiguation or non-standard speech. Additionally, using pre-trained models to fine-tune under-resourced languages can deepen the already rampant biases and their negative consequences for under-resourced languages and their respective communities. The research community should prioritize novel and under-resourced first approaches to leverage the unique characteristics of under-resourced languages that may not be present in high-resource languages. Furthermore, most under-resource language speakers are predominantly oral speakers, and a text-based machine translation is less accessible than a speech-based machine translation. The overwhelming majority of Bambara speakers who do not know how to read and/or write might not benefit right away from a text-based machine translation. Nevertheless, once speech synthesis systems for text-to-speech tasks become available, our work will be prevalent in deciding which state-of-the-art pre-trained models to consider in building applications for under-resource languages such as Bambara.

## References

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba O. Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Chukwuneke, Happy Buzaaba, Blessing Sibanda, Godson Kalipe, Jonathan Mukiibi, Salomon Kabongo, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Shamsuddeen Hassan Muhammad, Salomey Osei, Sokhar Samb, Tadesse Kebede Guge, and Pontus Stenetorp. 2024. Irokobench: A new benchmark for african languages in the age of large language models. *Preprint*, arXiv:2406.03368.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Ro-

man Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, et al. 2020. Findings of the 2020 conference on machine translation (wmt20). In *Proceedings of the Fifth Conference on Machine Translation*.

Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61.

Cathy Berthouzoz. 1999. Contextual resolution of global ambiguity for mt.

Charles Bird and Mamadou Kante. 1976. An ka bamanankan kalan: Intermediate bambara.

Charles S. Bird. 1966. *ASPECTS OF BAMBARA SYNTAX*. Ph.D. thesis. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-07-26.

Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. 2021. Transformers:" the end of history" for nlp? *arXiv preprint arXiv:2105.00813*.

Ousmane Daou, Satya Ranjan Dash, and Shantipriya Parida. 2024. Cross-lingual transfer learning for bambara leveraging resources from other languages. In *Empowering Low-Resource Languages With NLP Solutions*, pages 183–197. IGI Global.

Ousmane Daou and Sushree Sangita Mohanty. 2024. Cultural survival heritage of bambara language by using nlp. In *Applying AI-Based Tools and Technologies Towards Revitalization of Indigenous and Endangered Languages*, pages 173–182. Springer.

Klaudia Dombrowsky-Hahn. 2020. Valentin vydrin, cours de grammaire bambara (ouvrage en réalité augmentée). *Linguistique et langues africaines*, (6):141–146.

Sonja Ermisch. 2013. The structure of bambara. https://user.uni-frankfurt.de/~tezimmer/HP_FG-RelS/english/PDF/The%20Structure%20of%20Bambara_Handout_Budapest%20Mai%202013.pdf. Online; accessed 11 September 2024.

Christopher Ryan Green. 2010. *Prosodic phonology in Bamana (Bambara): Syllable complexity, metrical structure, and tone*. Ph.D. thesis, Indiana University.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. 2018. Universal neural machine translation for extremely low resource languages. *arXiv preprint arXiv:1802.05368*.

Basra Jehangir, Saravanan Radhakrishnan, and Rahul Agarwal. 2023. A survey on named entity recognition — datasets, tools, and methodologies. *Natural Language Processing Journal*, 3:100017.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, et al. 2020. Masakhane–machine translation for africa. *arXiv preprint arXiv:2003.11529*.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Amy Pu, Hyung Won Chung, Ankur P Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for mt. In *Proceedings of EMNLP*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Rochester Institute of Technology. 2024. Research computing services.

Allahsera Auguste Tapo, Bakary Coulibaly, Sébastien Diarra, Christopher Homan, Julia Kreutzer, Sarah Luger, Arthur Nagashima, Marcos Zampieri, and Michael Leventhal. 2020. Neural machine translation for extremely low-resource African languages: A case study on Bambara. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*.

The Dokotoro Project. Bambara and french language versions of "where there is no doctor: A village health care handbook". https://dokotoro.org/. Online: Accessed October 10, 2024.

Atnafu Lambebo Tonja, Bonaventure F. P. Dossou, Jessica Ojo, Jenalea Rajab, Fadel Thior, Eric Peter Wairagala, Anuoluwapo Aremu, Pelonomi Moiloa, Jade Abbott, Vukosi Marivate, and Benjamin Rosman. 2024. Inkubalm: A small language model for low-resource african languages. *Preprint*, arXiv:2408.17024.

Valentin Vydrin. 2009. On the problem of the proto-mande homeland. *Journal of language relationship*, 1:107–142.

Valentin Vydrin. 2013. Bamana reference corpus (brc). *Procedia-Social and Behavioral Sciences*, 95:75–80.

Valentin Vydrin. 2014. Projet des corpus écrits des langues manding: le bambara, le maninka. In *Traitement Automatique du Langage Naturel 2014*.

Valentin Vydrin. 2018. Mande languages. In *Oxford Research Encyclopedia of Linguistics*.

Valentin Vydrin, Kirill Maslinsky, Jean-Jacques Méric, and A Rovenchak. 2011. Corpus bambara de référence.

Valentin Vydrin, Andrij Rovenchak, and Kirill Maslinsky. 2016. Maninka reference corpus: A presentation. In *TALAf 2016: Traitement automatique des langues africaines (écrit et parole). Atelier JEP-TALN-RECITAL 2016-Paris le*.

Valentin Vydrine. 2015. *Manding-English Dictionary: Maninka, Bamana Vol. 1.*, volume 1. MeaBooks.

Claire Weeks. 2021. Machine translation for low-resource languages: a community-based participatory approach.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.