Bayelemabaga: Creating Resources for Bambara NLP

Anonymous ACL submission

Abstract

In under-resourced language technology settings, the problem is often not only the amount of data available, but also the quality, and in ways that are entirely foreign to high-resourced languages. For instance, most of the world's languages have only recently acquired writing systems, resulting in multiple writing systems competing for dominance or, within a single writing system, non-standardized spelling. In this case study, we focus on the impact of manual data cleaning on the performance of learning machine translation models on Bambara, the vehicular language of Mali, and introduce the largest curated dataset for multilingual translation. We finetune 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 our newly curated dataset, Bayelemabaga, and evaluate on four Bambara-French language pair datasets (three existing datasets and the test set of Bayelemabaga). We show that our newly aligned and curated multilingual dataset enhances the translation quality of all studied models using the BLEU, CHRF++, and AfriCOMET evaluation metrics.

1 Introduction

Driven by the availability of massive digitized data stores as much as by advancements in neural architectures (Chernyavskiy et al., 2021), state-of-the-art natural language processing (NLP) tools are widely adopted by the world's high-resource languages (e.g., English, French) to solve complex tasks in domains 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), etc.

Yet the vast majority of the world's languages, and by extension the people who speak these as their primary or sole language, lack the digitized data resources needed to support MT systems (Weeks, 2021) such as Google Translate. These under-resourced languages have yet to benefit from recent advances because they lack the large volumes of translated texts needed to drive neural machine translation (Gu et al., 2018). Moreover, the available data is often noisy and diverse in ways that are less common in higher-resource language, 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, e.g., Adjami (Arabic-based), Latin and N'ko scripts. However, as a historically predominately oral language, a majority of Bambara speakers have never been taught to read or write the standard form of the language. As a result, available resources don't always fit a common writing system (e.g., systems developed during colonization), and lack standard orthography and ways to express features such as tonality that are not present 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 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.

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's[2] Corpus Bambara de

[1] https://en.wikipedia.org/wiki/Mande_languages
[2] http://www.inalco.fr/en

Reference,[3] aligned collected sentences in both languages, investigated their morphological structure, and curated the content to ensure adequacy for machine translation. We then evaluate the adequacy of Bayelemabaga by answering the following research questions: (1) How does the quality of our curated dataset compare to the raw data? (2) What is the impact of our dataset on improving translation results compared to existing models fine-tuned on the currently scarce data? (3) 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 resource for training, and adaptable 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 the language, there is a need for 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.

Educational materials for learning Bambara are relatively scarce, particularly in comparison to more widely taught languages like high-resource languages such as French or English. However, some resources do exist, primarily in the form of 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 of academic or linguistics research settings. There is a need for more modern, 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 they 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 promote 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 (Farhad 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.

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 (e.g., 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 main 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

[3] http://cormand.huma-num.fr/

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 linguistics work over the decades, 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 such as HuggingFace and GitHub [4].

3.1 Data Collection

The Bayelemabaga dataset was curated from 231 data sources ranging from informal content (short stories, blog posts), and 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 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

Our data alignment and curation approach follows a three-step strategy described as follows.

In the first step, we invited linguists and language experts to group collected Bambara and French sentences into two 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".

The second step consisted of an automated alignment of the sentences wherein we parsed each map-

[4] anonymous

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.
Train	37,580	15	531,501	37,975
Dev	4,698	16	73,727	8,388
Test	4,698	10	44,278	5,516

ping to generate a JSON file that combined each French sentence with its equivalent Bambara translation into one dictionary object. During this automated alignment step, we skipped all instances of "$nTAB - 1$" and "$-1TABm$" ensuring every sentence has a reference translation.

After completing the sentence alignment step, we applied a curation procedure using Python's regular expression (re) library to clean both the Bambara and French files. This process involved removing unwanted elements such as HTML/XML tags, non-printable characters, and orphan symbols. Additionally, we eliminated any newline characters within aligned pairs of French and Bambara sentences to ensure data consistency.

3.3 Data Partitioning

To ensure models trained with our dataset are robust and generalize well to new data, Bayelemabaga was divided into training (Train), validation (Dev), and evaluation (Test) 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 the number of unique tokens.

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 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.

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 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 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).

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

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

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 gener-

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

ation 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.

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 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.

4.3 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 pre-trained version of M2M-100 scores 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 and curation process. Table 5 reports the comparison results between models fine-tuned on the collected data (Raw) and Bayelemabaga. These results highlight that considering the Bambara language nuances during our alignment and curation of the collected dataset improves translation quality across all models evaluated on the test set of the 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 a model fine-tuned on datasets with several uninformative and meaningless characters and words is less likely to generate logical and structured translations of curated data and vice-versa.

Bi-directional Machine Translation: We next contrast the contributions of Bayelemabaga to machine translation with encoder-decoder and decoder-only transformer-based models on bi-directional translation tasks. These analyses allow us to assess the usefulness of our data for both Bambara and French speakers and our ability to contribute to the landscape of large decoder-only language models. The training and evaluation datasets combine corpora from the Dictionary, Medical, News, and Bayelemabaga datasets, yielding a train set of 42,526 sentences, a validation set of 6,474 sentences, and an evaluation set of 6,616 sentences. The results in Table 6 demonstrate that encoder-decoder models outperform decoder-only models in machine translation. The performance gap is 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 boost offered by presenting a few example

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 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 finetuned on the training set of all datasets (Dict., 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

reference-translation pairs to decoder models. Figure 1 illustrates the results of different numbers of examples (0, 1, 3, and 5). Compared to CHRF++, the BLEU evaluation metric shows high variability across different shots, suggesting inconsistent translation quality. CHRF++ better captures partial matches and morphological features of the Bambara language 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 are pre-trained on massive multilingual datasets, including languages that expose patterns similar to Bambara. Combined with our carefully curated dataset, the decoder models learned a compact and 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 19 translations from each of the seven models. The samples were selected randomly across all translated sentences, but the same sentences were evaluated for all models. To ensure fairness, we recruited two evaluators and presented them with the translations without knowledge of the model used to generate them. As metrics, we evaluated the quality of the translation ("good" or "bad"), problems associated with "fluency", adequacy", and "understanding", and introduced a score of one to five to rank the model's overall performance, five being the best score. More precisely, we indicated fluency as a problem when the model's output does not read easily. We made adequacy an issue for outputs that do not contain the ideal word choice in conveying the desired meaning and marked understanding as a problem when the output does not make sense to the Bambara speaker reading it.

Table 7 summarized our human evaluation results. We found that the NLLB-200 model performs best, with an average score of 4.58. All evaluated translations were "good," but 1 sample had a fluency problem and 3 had an adequacy problem.

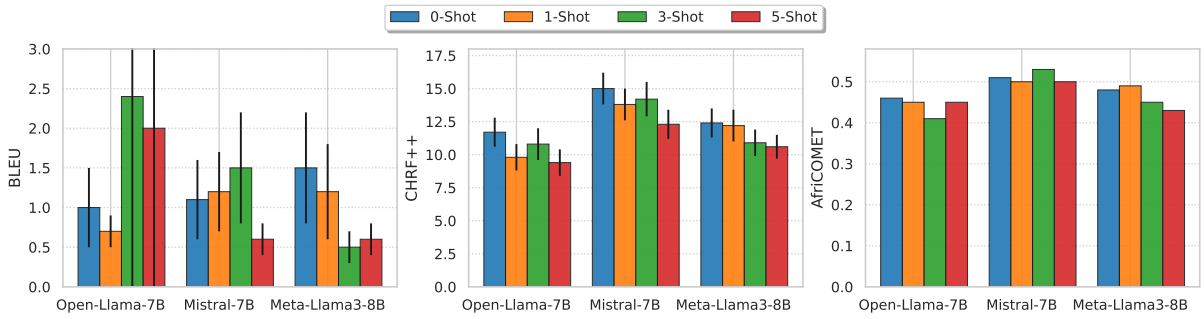

Figure 1: Evaluation of decoder models on translations from French to Bambara with 0-shot, 1-shot, and few-shots.

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 (fluency/adequacy/understanding)
MBART	4.21	16/3	1/3/2
MT5	3.74	12/7	2/3/4
NLLB-200	**4.58**	**19/0**	**1/3/0**
M2M-100	3.95	15/4	1/4/3
MISTRAL	1.11	0/19	0/0/19
Open-Llama-7B	1	0/19	0/0/19
META-Llama3-8B	1	0/19	0/0/19

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

Table 8 highlights a few examples of translation that were evaluated and their corresponding 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 into "A bolo" which means "in its hand", which does not align with the correct translation. Another example is 3, a translation of "eau froide" ("cold water" in English) into "ji sumalen" by the M2M-100 model. We notice an additional phrase "len" compared to the reference, but both can equally be used as good translations.

The human evaluation scores support the machine translation results as presented in Table 6 where the best models are encoder-decoder architectures, with NLLB-200 generating the best translations. The worst score from the decoder-only models is also supported as none of their generated translations ranked as good or was understood by native speakers. While they present sentences that may have a meaning in the target language, they fail to properly translate the source text, which is the reason behind the low scores.

5 Conclusion

Creating resources for under-resourced languages is crucial for preserving and promoting their important cultural heritage and Indigenous know-how. Additionally, the creation of digital resources, e.g., corpora, and machine learning tools, has the potential to facilitate their integration into modern technologies, thus promoting their use in various domains. In this work, we explored the effect of curating data in machine translation compared to just utilizing it in its raw, unclean state. Taking into account the garbage in, and garbage out assumption, coupled with our observations, we observe in most of our experiments that efforts put into curating a dataset have the potential to yield dividends, by improving the outcome of the models' output.

In future work, we plan to conduct a human evaluation of both outputs from models fine-tuned on curated and raw data to see how human insights compare to automatic metrics. We also plan on investigating further for primarily oral languages (POLs) in speech modality whether our observations in text modality apply in the speech realm.

6 Limitations

Two major issues for machine translation are ambiguity and non-standard speech (Berthouzoz, 1999; Koehn and Knowles, 2017). In this work, we do not directly address disambiguation or non-standard speech. Utilizing pre-trained models to fine-tune under-resourced languages has the potential to deepen the already rampant biases and their negative consequences for under-resourced languages and their respective communities. Novel and under-resourced first approaches should be prioritized 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 to them than a speech-based machine translation. We anticipate that 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 (aka text-to-speech) systems become available, work, like ours will be handy in deciding which SOTA pre-trained models to consider in building applications for low-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.

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.

Akhbardeh Farhad, Arkhangorodsky Arkady, Biesialska Magdalena, Bojar Ondřej, Chatterjee Rajen, Chaudhary Vishrav, Marta R Costa-jussa, España-Bonet Cristina, Fan Angela, Federmann Christian, et al. 2021. Findings of the 2021 conference on machine translation (wmt21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88. Association for Computational Linguistics.

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*.