Bayelemabaga: Creating Resources for Bambara NLP

Anonymous ACL submission

Abstract

In low-resource 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, many extreme low-resource languages have only recently acquired writing systems. This may result in multiple writing systems competing for dominance or, within a single writing system, non-standardized spelling. Translating to and from low-resource languages is a challenge for machine translation (MT) systems due to a lack of suitable parallel data. In this case study, we focus on the impact of manual data cleaning on the performance of learning machine translation models. We focus on Bambara, the vehicular language of Mali, and introduce the largest curated dataset for multilingual translation. We finetune six commonly used transformer-based language models, i.e., AfriMBART, AfriMT5, AfriM2M100, Mistral, Open-Llama-7B, and Meta-Llama3-8B on three existing Bambara-French language pair datasets and our curated dataset. We show that our new aligned and curated multilingual dataset enhances the translation quality of all studied models using the BLEU, CHRF++, and AfriCOMET evaluation metrics.

1 Introduction

State-of-the-art natural language processing (NLP) tools are available and utilized by the world's highly-resourced languages (e.g. English, French, etc.). In domains they excel at machine translation (MT) with 30+ (BLEU) scores (Wu et al., 2016), name entity recognition (NER), automatic speech recognition (ASR), etc. These advancements were enabled by abundant digitized resources available, along with advancements in neural architectures (Chernyavskiy et al., 2021).

Unfortunately, the vast majority of the world's languages, spoken by a majority of the world's population lack digitized resources and are missing on existing publicly available MT systems Weeks (2021) e.g., Google Translate. These under-resourced languages have yet to benefit from these advances, because they lack the large volumes of translated texts needed to drive neural machine translation (Gu et al., 2018). Beyond a paucity of data, the data available is noisy and extremely heterogeneous, with non-standardized spelling, accenting, marking, multiple scripts, code-switching etc.

For example, Bambara is a tonal language with a rich morphology. Over the years, several competing writing systems have been developed, 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. Many are incapable of reading or writing the language at all. Among the complications of working with under-resourced languages is that they don't always fit the writing systems imposed on them through colonization. So that means they lack standard orthographies, but also lack standard ways to express features such as tonality that are not present in colonial scripts. This in turn has led to competing writing systems. In the case of the Mande family of languages, Adjami script (Arabic based), Latin script, and N'ko script. The standardization of words and the coinage of new ones are still works in progress; this poses challenges to automated text processing.

We have seen a number of initiatives in this direction, such as Masakhane for African languages Orife et al. (2020), the annual Conference on Machine Translation (WMT) that increasingly includes low-resource languages in their popular machine translation (MT) competitions (Barrault et al., 2019, 2020), and AfricaNLP, a dedicated workshop focused on African languages. To further address the scarcity of data for machine translation in low-resource 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 [1]'s Corpus Bambara de Reference [2], aligned collected sentences in both languages, investigated their morphological structure, and curated the content to ensure adequacy for machine translation. We thoroughly evaluate the adequacy of the *Bayelemabaga* dataset focusing on three key research questions: (1) How does the quality of our curated dataset compare to uncleaned 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 in the target language, perform when evaluated using our dataset? The *Bayelemabaga* dataset is intended to improve the performance of translation models for the Bambara language by providing a richer resource for training, and adaptable to other natural language processing tasks.

2 Background and Motivation

2.1 The Mande Language Family

The Mande language family consists of several languages spoken by 30–40 million people across the African continent. Among them, we explore Bambara 2.1.1 language in this paper.

2.1.1 The Bambara Language

Described in (Tapo et al., 2020) as being tonal (e.g., different words with different inflections convey different meaning) with a rich morphology, and a number of competing writing systems. Bambara is the mother tongue to Bambara people in Mali, across the African continent, and beyond. It has two underlying tones (high and low). It is closely related to other Manding languages such as Maninka (or Malinke), Jula (or Dyula), Mandinka, etc. (sometimes considered as dialects of one language). Diacritics in Bambara are tonal markers. Bambara is written in many writing scripts such as adjami, latin, and N'ko. In its latin writing script has 27 letters without q, v, x.

Examples. Below are some examples of Bambara usage.
French: les remèdes maison utiles
Bambara: "farafinfura minnu bɛ se ka bana furakɛ"
English: the useful homemade medication

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

3 Related Work

Several linguistic studies have been conducted on the Bambara language, providing valuable insights into its structure (**?**), 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.

Furthermore, in the digital age, the availability of online resources and technological tools can greatly enhance language learning and preservation efforts. However, Bambara has a relatively limited presence in the digital realm.

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), 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.

Despite being a widely spoken language, Bambara faces challenges in terms of preservation and promotion, particularly in the face of the dominance of colonial languages like French and the increasing influence of globalization.

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 efforts involving linguists, educators, technology experts, and community stakeholders to curate decades of linguistic data from varying sources including books, periodical, news etc. for machine learning, including machine translation.

4 The Bayelemabaga Dataset

We created a parallel text dataset for the dialect continuum of Manding languages spoken in West Africa by approximately 40 million people. We collected 46,976 parallel sentences, ready to be utilized, and already available for download on HuggingFace as well as on GitHub, and referenced on Lanfrica's data store. In order to enable standard orthographically sound writing for Bambara, we held a workshop where all the leading Bambara linguists gave the Bambara language a complete scientific orthographic system.

The Bambara dataset contains data from Dokotoro, the Bible, SIL Dictionary Sentences, and some data from the *Corpus Bambara de Référence* (Vydrin et al., 2011). The sentences in Bambara are in the Latin script.

4.1 Data Collection

The Bayelemabaga dataset is a curation of 46,976 Bambara and French sentences, curated from 231 data sources, varying from periodicals, books, short stories, blog posts, part of the Bible and the Quran from INALCO [3]'s*Corpus Bambara de Référence* (Vydrin et al., 2011).

There are four types of text in the initial parallel corpus. A type is like a stage in which the data is in, such as annotated or not, disambiguated or not, adjusted or not. The types are as follows:

- Type 1: Non annotated, non disambiguated, plus its French translation.
- Type 2: Annotated Bambara, with its French translation.
- Type 3: Annotated Bambara, plus two French translations, with the second French translation adjusted.
- Type 4: Annotated Bambara, and its adjusted French translation.

Type 3 being the most complete in terms of having the annotated Bambara, the original French translation, and the adjusted French translation.

4.2 Preparation

For a given data source, three files are created using the name of the data source. The first contains Bambara text , the second contains French text , and the third contains information to map Bambara text to French text. This mapping information is represented as numbers, separated by "TAB".

The mapping information is represented as "$nTABm$", where n is the number of the sentence(s) in Bambara, m is the number of the sentence(s) in French, and TAB is the delimiter separating n from m.

Sentence Alignment Utilizing the ".prl" files, we wrote a python script that parses them one at a time from a bash script, and make sense of the mapping, then it writes the aligned Bambara/French into a JSON file.

Preprocessing During alignment, we skipped all "$nTAB-1$" and "$-1TABm$", "-1" stands for "no matching" French or Bambara respectively. After completing the alignment step at sentence level. We utilized python's re library to remove tags, non-printable characters, etc. in both Bambara and French files.

[3] http://www.inalco.fr/en

Table 1: Overview of the datasets. Dictionary, Medical, News, and Lacuna 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 2: 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

Parallel Data. The number of parallel sentences is found in Table 1.

The final dataset contains 46,976 Bambara-French parallel sentences. We split it into train, dev, and test sets with 80%, 10%, and 10% respectively.

5 Experiments

We evaluate the quality of the Lacuna 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.

5.1 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.

5.2 Methods

Our evaluation focuses on two main classes of transformer-based language models, i.e., encoder-decoder and decoder-only models, instrumented for the machine translation to generate text in Bambara or French depending on the evaluated source language.

Table 3: Performance of MT models fine-tuned on raw and cleaned versions of the Lacuna Dataset and evaluated on the test sets of raw and cleaned versions.

Models	Lacuna-Raw		
	BLEU	CHRF++	AFRICOMET
AfriMBART-Raw	**7.8 ± 1.9**	13.6 ± 1.5	0.18
AfriMBART-Clean	0.9 ± 0.7	**16.2 ± 3.9**	**0.27**
AfriMT5-Raw	**13.3 ± 5.5**	**11.5 ± 1.6**	**0.11**
AfriMT5-Clean	2.0 ± 1.4	6.4 ± 2.9	0.05
AfriM2M100-Raw	**8.1 ± 2.2**	15.6 ± 2.2	0.25
AfriM2M100-Clean	1.0 ± 0.7	**16.4 ± 3.9**	**0.26**
	Lacuna-Clean		
	BLEU	CHRF++	AFRICOMET
AfriMBART-Raw	0.4 ± 0.2	10.6 ± 1.8	0.35
AfriMBART-Clean	**5.0 ± 1.7**	**24.6 ± 1.9**	**0.50**
AfriMT5-Raw	0.3 ± 0.1	5.2 ± 0.8	0.10
AfriMT5-Clean	**1.6 ± 0.7**	**16.6 ± 1.8**	**0.37**
AfriM2M100-Raw	5.0 ± 1.9	23.6 ± 1.8	**0.55**
AfriM2M100-Clean	**6.4 ± 2.0**	**25.8 ± 2.0**	0.53

5.2.1 Data

Our analyses focus on four different datasets, i.e., (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 (§ 4). We also use the version of the dataset before curation to assess the quality of our curation and alignment effort (§ 5.3.1.

For each dataset, we randomly split the data into training (80%), validation (10%), and test (10%) sets. Table 1 provides an overview of each dataset.

5.2.2 Models

Encoder-Decoder We experiment with three encoder-decoder models finetuned with datasets from 16 African languages by Adelani et al. (Adelani et al., 2022): (1) *AfriMBART* [4]. a fine-tuned MBART model is tailored for sequence-to-sequence multilingual tasks, enhancing its translation

[4] https://huggingface.co/masakhane/afrimbart_fr_bam_news

Table 4: Summary of the evaluation performance of pre-trained and finetuned LLMs on evaluation datasets from diverse domains.

	Dict. (2K)	Medical (4K)	News (6K)	Lacuna (47K)
		BLEU		
AfriMBART	0.1 ± 0.1	1.3 ± 1.3	1.3 ± 1.1	0.3 ± 0.1
AfriMBART-Lacuna	**20.6 ± 6.3**	**4.8 ± 1.8**	**14.1 ± 2.7**	**5.0 ± 1.7**
AfriMT5	0.4 ± 0.3	0.9 ± 1.1	2.4 ± 1.3	0.5 ± 0.2
AfriMT5-Lacuna	**3.4 ± 1.6**	**3.9 ± 2.0**	**2.7 ± 1.1**	**1.6 ± 0.7**
AfriM2M100	**27.8 ± 7.4**	**7.6 ± 2.2**	**18.0 ± 3.1**	4.6 ± 1.6
AfriM2M100-Lacuna	25.0 ± 6.3	7.2 ± 1.9	11.6 ± 2.0	**6.4 ± 2.0**
		CHRF++		
AfriMBART	2.0 ± 0.6	6.1 ± 1.3	10.8 ± 2.0	5.4 ± 1.6
AfriMBART-Lacuna	**37.4 ± 5.0**	**20.5 ± 1.6**	**39.2 ± 2.4**	**24.6 ± 1.9**
AfriMT5	7.8 ± 2.0	4.7 ± 0.6	12.5 ± 2.0	7.5 ± 1.0
AfriMT5-Lacuna	**17.0 ± 3.2**	**11.0 ± 1.6**	**13.9 ± 1.7**	**16.6 ± 1.8**
AfriM2M100	**48.2 ± 5.7**	**24.5 ± 1.9**	**43.6 ± 2.7**	24.3 ± 1.8
AfriM2M100-Lacuna	47.7 (± 5.2)	24.4 ± 1.7	35.7 ± 2.3	**25.8 ± 2.0**
		AFRICOMET		
AfriMBART	0.33	0.22	0.27	0.30
AfriMBART-Lacuna	**0.61**	**0.42**	**0.53**	**0.50**
AfriMT5	0.22	0.1	0.23	0.22
AfriMT5-Lacuna	**0.37**	**0.22**	**0.30**	**0.37**
AfriM2M100	0.66	0.47	**0.57**	**0.55**
AfriM2M100-Lacuna	0.66	0.47	0.54	0.53

and text generation capabilities across various languages. (2) *AfriMT5* [5]. a multilingual variant of the T5 model employed for various tasks including translation, summarization, and question-answering tasks. (3) *AfriM2M100* [6]. a multilingual MT model designed to handle many-to-many language translations between any pair of 100 languages.

Decoder-only We further explore the quality of *Bayelemabaga* for MT tasks using emerging transformer-based language model architectures that only feature decoders. Although the type of model is aimed at general-purpose language generation tasks, several existing efforts have adapted them for specific tasks, including machine translation. We identified the following three open-source models available for free for our research: (1) *Open-Llama-7B* [7]. an open-source adaptation of the LLaMA model, aimed at general-purpose language understanding and generation, can be fine-tuned for MT tasks. (2) *Mistral-7B* [8] 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* [9] a language model developed by Meta AI and optimized for various NLP tasks.

5.2.3 Hyperparameters

We present the hyperparameters for finetuning and evaluating selected models with the *Bayelemabaga* dataset in Table 2. Most hyperparameters for encoder-decoder LLMs follow existing work on machine translation in Bambara (Adelani et al., 2022). The default hyperparameters on Huggingface Transformers are adopted if not included in Table 2.

5.2.4 Metrics

We compare the quality of different MT systems using widely-known n-gram matching evaluation metrics, ScareBLEU (BLEU) (Post, 2018) and CHRF++ (Popović, 2015). We also use AfriCOMET (Pu et al., 2021), a learned COMET

[5] https://huggingface.co/masakhane/afrimt5_fr_bam_news

[6] https://huggingface.co/masakhane/m2m100_418M_fr_bam_news

[7] https://huggingface.co/openlm-research/open_llama_7b

[8] https://huggingface.co/mistralai/Mistral-7B-v0.3

[9] https://huggingface.co/meta-llama/Meta-Llama-3-8B

Table 5: Summary of the performance of encoder-decoder and decoder-only LLMs finetuned on all datasets (Dict., Medical, News, and Lacuna) using different combinations of source and target languages. The best results for decoder-only models are obtained with a 5-shot inference.

Models	$fr \rightarrow bam$			$bam \rightarrow fr$		
	BLEU	CHRF++	AFRICOMET	BLEU	CHRF++	AFRICOMET
Encoder-Decoder						
AfriMBART	29.5 ± 7.8	$\mathbf{47.9 \pm 5.6}$	**0.66**	$\mathbf{9.4 \pm 4.0}$	$\mathbf{30.4 \pm 4.2}$	0.33
AfriMT5	5.8 ± 2.6	21.8 ± 3.5	0.44	0.2 ± 0.1	7.1 ± 1.1	0.54
AfriM2M100	$\mathbf{33.1 \pm 7.7}$	51.8 ± 5.6	**0.66**	1.4 ± 1.6	7.6 ± 1.0	**0.57**
Decoder-only						
Mistral-7B	1.2 ± 0.5	11.6 ± 2.1	0.53	1.3 ± 1.1	11.3 ± 1.7	0.31
Open-Llama-7B	2.1 ± 1.3	10.0 ± 1.9	0.51	1.2 ± 1.6	8.1 ± 1.2	0.24
Meta-Llama3-8B	0.6 ± 0.3	9.5 ± 1.5	0.51	0.8 ± 0.7	10.7 ± 1.9	0.28

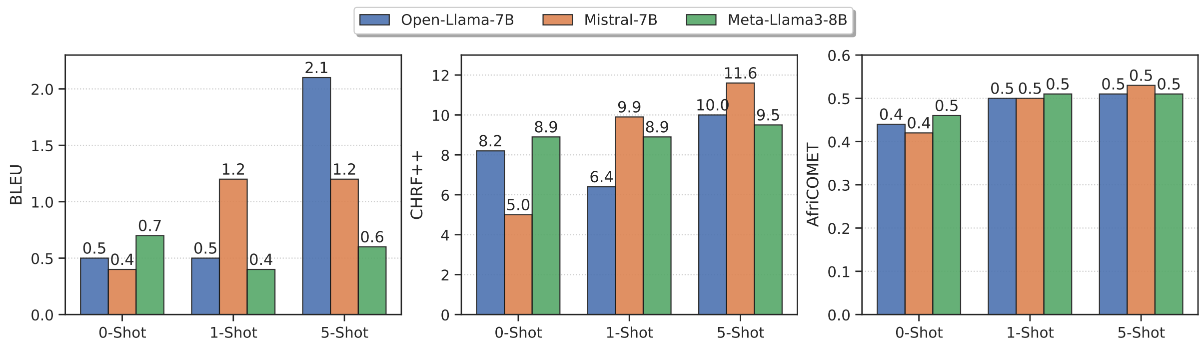

Figure 1: Comparison of performance of decoder-only LLMs on machine translation tasks from French to Bambara w.r.t the numbers of observed examples

metric for MT evaluation covering 13 topologically different African languages.

5.3 Results

5.3.1 Automatic Evaluation

Alignment and Curation Quality Assessment: Our first set of experiments evaluates the quality of our data alignment and curation. We finetune existing state-of-the-art language models on the collected data (Raw) and the final curated *Bayelemabaga* datasets and compare their evaluation scores. Table 3 reports the comparison results. We observe that fine-tuning MT models on the *Bayelemabaga* dataset enhances the quality of the translations, yielding up to +4.4, +14, and +0.27 in gains respectively on BLEU, CHRF++, and AfriCOMET scores, when evaluated on the test set of the curated data. Interestingly, the fine-tuning and evaluation of MT models on the raw data shows better scores than fine-tuning on the clean data and evaluating with the raw data. This behavior is expected because a model finetuned on datasets with several uninformative and meaningless characters and words cannot generate logical and structured translations of curated data. These results highlight the correctness of our evaluation and the quality of the *Bayelemabaga* dataset.

To further investigate the quality of *Bayelemabaga*, we compare our finetuned models with the pre-trained versions across different datasets. As reported in Table 4, fine-tuning with the *Bayelemabaga* dataset enhances the quality of generated translations. The pre-trained version of AfriM2M100 scores slightly better than the finetuned version on average on the dictionary, medical, and news data, but the variance indicates that both systems are comparable. Additionally, these three datasets contain fewer sentences and were originally included in the data used to pre-train the models, presenting risks of overfitting.

Reverse Translation Evaluation: We next evaluate the quality of MT models, initially fine-tuned for translation from French to Bambara, by assessing their performance on translation from Bambara to French. We conducted this evaluation using encoder-decoder and decoder-only models to compare their effectiveness in bidirectional translation tasks. The results in Table 5 demonstrate that encoder-decoder models outperform decoder-only models in machine translation because encoder-decoder architectures have distinct encoding and decoding phases, better equipped to handle the complexity of capturing and translating context from the source to the target language. However, we found that decoder-only models maintain consistent results when translating from French or Bambara due to their ability to leverage language modeling capabilities, e.g., contextual understanding, autoregressive generation capabilities, and bidirectional language pattern recognition.

Impact of zero-shot and few-shot translations on Decoder Models: We also investigate the performance boost offered by presenting a few example reference-translation pairs to decoder models before a translation request. We experiment with different numbers of examples (0-shot, 1-shot, and 5-shot) for decoder-only models. Figure 1 shows that increasing the number of examples yields a better score due to the model's ability to use the provided examples to understand the specific translation patterns and nuances. These findings align with prior work on machine translations with decoder-only models, which has similarly observed that providing more examples leads to improved performance for African languages (Adelani et al., 2024).

6 Conclusion

Creating resources for low-resourced languages is crucial for preserving and promoting their important cultural heritage as well as indigenous know-hows. Additionally, the creation of digital resources, such as corpora, machine learning tools, has the potential to facilitate their integration into modern technologies, thus promote 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, garbage out assumption, coupled with our observations, we observe in most of our experiments that efforts put in to curating a dataset has the potential to yield dividend, by improving outcome of the models' output.

As future work, we plan conducting human evaluation of both outputs from models fine-tuned on curated and raw data to see how human's insights compare to automatic metrics. Additionally, we plan on investigating further for primarily oral languages (POLs) in speech modality weather our observations in text modality apply in the speech realm.

7 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 nor non-standard speech.

Utilizing pretrained models to fine tune on low-resourced languages has the potential to deepen the already rampant biases and their negatives consequences for low-resource languages and their communities.

Novel, and low-resourced first approaches should be prioritized to leverage unique characteristics of low-resource languages that may not be present in high-resource languages.

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.

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*.

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*.

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*.

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*.