A Few Thousand Translations Go A Long Way!
Leveraging Pre-trained Models for African News Translation
David Ifeoluwa Adelani∗, Jesujoba Oluwadara Alabi∗, Angela Fan∗, Julia Kreutzer∗,
Xiaoyu Shen†, Machel Reid†, Dana Ruiter†, Dietrich Klakow†, Peter Nabende‡,
Ernie Chang‡, Tajudeen R. Gwadabe§, Freshia Sackey§, Bonaventure F. P. Dossou**,
Chris Chinenye Emeze††, Colin Leong††, Michael Beukman††,
Shamsuddeen H. Muhammad††*, Guyo D. Jarso*, Oreen Yousuf††*, Andre N. Rubungo††*,
Gilles Hacheme‡‡, Eric P. Wairagala‡‡, Muhammad U. Nasir‡‡*, Benjamin A. Ajibade‡‡,
Tunde Oluwaseyi Ajayi‡‡, Yvonne Wambui Gita‡, Jade Abbott‡, Mohamed Ahmed‡‡*,
Millicent Ochieng‡‡*, Anuoluwapo Aremu‡, Perez Ogayo‡‡*, Jonathan Mukiibi‡‡*,
Fatoumata Ouoba Kabore‡, Godson Koffi Kalipe‡, Derguene Mbaye‡‡‡,
Allahsera Auguste Tapo‡‡‡, Victoire M. Koagne‡, Edwin Munkoh-Buabeng‡,
Valencia Wagner†††, Idris Abdulmumin†††, Ayodele Awokoya†††, Happy Buzaaba‡,
Blessing Sibanda†‡‡, Andiswa Bukula†‡‡, Sam Manthulu†‡‡

*Masakhane NLP, 1 Saarland University, Germany, 2Inria, France, 3Meta AI, 4Google Research, 5Amazon Alexa AI,
6The University of Tokyo, Japan, 7Makerere University, Kampala, Uganda, 8UCAS, China, 9JCUAT, Kenya,
10 Jacobs University, Germany, 11TUM, Germany, 12University of Dayton, USA, 13University of the Witwatersrand, South Africa,
14LIAAD-INESC TEC, Porto, Portugal, 15Uppsala University, Sweden, 16UPC, Spain, 17UiHinno18Omninor AI,
19Microsoft Africa Research Institute, Kenya 20CMU, USA, 21Baamtu, 22RIT, USA, 23SPU, South Africa,
24ABU, Nigeria, 25UI Ibadan, Nigeria, 26NUST, Namibia 27SADILaR, South Africa, 28University of Malawi, Malawi

Abstract
Recent advances in the pre-training of language models leverage large-scale datasets to create
multilingual models. However, low-resource languages are mostly left out in these datasets.
This is primarily because many widely spoken languages are not well represented on the web
and therefore excluded from the large-scale crawls used to create datasets. Furthermore,
downstream users of these models are restricted to the selection of languages originally
chosen for pre-training. This work investigates how to optimally leverage existing pre-trained
models to create low-resource translation systems for 16 African languages. We focus on
two questions: 1) How can pre-trained models be used for languages not included in the
initial pre-training? and 2) How can the resulting translation models effectively transfer
to new domains? To answer these questions, we create a new African news corpus covering
16 languages, of which eight languages are not part of any existing evaluation dataset.
We demonstrate that the most effective strategy for transferring both to additional languages
and to additional domains is to fine-tune large pre-trained models on small quantities of high-
quality translation data.

1 Introduction
Enormous efforts have been invested in making language and translation models more multilingual
while leveraging the maximal amount of data for training, most prominently large crawls of mono-
linguial and parallel data from the web (El-Kishky et al., 2020; Schwenk et al., 2021b;a; Xue et al.,
2021b). The resulting models are now capable of translating between hundreds of languages, including
language pairs that in isolation do not have large collections of parallel data (Tang et al., 2020;
Xue et al., 2021a; Fan et al., 2021b). For example, M2M-100 (Goyal et al., 2021) can translate (with
low accuracy) between Hausa and Yorùbá, two of the most widely spoken languages in Nigeria, even
though there is barely any parallel data available for training. For languages that are not included in the
set of training languages, the model would have no knowledge on how to generate translations. Does
this mean there is no hope for languages that do not have large presence on the web and are therefore
not included in these pre-trained models?

We investigate how large-scale pre-trained models can be leveraged for the translation of unseen
low-resource languages and domains. We address this question by studying 16 African languages that
are largely underrepresented in NLP research (Joshi et al., 2020; ∗ et al., 2020) and further have little
to no training data available (§3). These languages provide an ideal testbed for two challenging knowl-
edge transfer tasks: (1) How can pre-trained models create translations for languages unseen at train-
ing time? and (2) Since training data may only exist in single domain (i.e. religious texts), how can a model be trained in one domain and translate another effectively at test time?

These questions are extremely relevant for our chosen languages because they all have millions of native speakers and a massive need for translation technologies. For example, news concerning the African continent are almost exclusively published in English, French, or Arabic, and thereby inaccessible for speakers of only native African languages. This creates a bottleneck for information transmission, which becomes even more critical in times of crises (Öktem et al., 2020; Anastasopoulos et al., 2020; Öktem et al., 2021). Furthermore, the task of translating news has historically played a central role in translation research, e.g. in shared tasks since 2008 (Callison-Burch et al., 2008) and as a test for determining human parity (Hassan et al., 2018; Läubli et al., 2018; Toral et al., 2018). To spur the development of dedicated news translation models for Africa, we construct a benchmark of news translation for translating between 16 native African languages and English or French (§4).

This allows us to compare three approaches to leveraging large-scale multilingual models for the translation of previously unseen languages: (1) zero-shot transfer, (2) continual pre-training on monolingual data, and (3) multi-domain fine-tuning on parallel data (§5). We find that fine-tuning pre-trained models on a few thousand sentences of high quality bitext is remarkably effective, and can be further augmented with continual pre-training on African languages and fine-tuning on news domain data (§6). Our contributions are the following:

1. We create a new African news corpus for machine translation (following principles of participatory research ∀ et al. (2020)) covering 16 African languages.

2. We adapt several multilingual pre-trained models (MT5, ByT5, mBART, M2M-100) to these largely unseen languages, and evaluate their quality on news translation.

3. We quantify the effectiveness of small in-domain translation sets by comparing domain transfer effects and comparing fine-tuning strategies.

We find that having a targeted collection of translations is surprisingly effective, showcasing the power of local knowledge in so-called “zero-resource” scenarios (Bird, 2020). This paints a promising picture for the development of NLP technology for understudied languages: being able to customize these models for new language of interest with as little as 2k sentences and a few fine-tuning steps, MT developers and users from any language community are less dependent on choices and monetary interest of industry powerhouses from the Global North (Paullada, 2020).

2 Related Work

African MT Datasets. One of the major challenges of developing MT models for African languages is lack of data. There are many attempts to automatically crawl and align sentences from the web (Schwenk et al., 2021a,b). Nevertheless, the resulting corpora for many African languages are typically small and of poor quality (Kreutzer et al., 2021). Other cleaner parallel sources are mostly from religious sources, like the Bible covering over 1600 languages (McCarthy et al., 2020) and JW300 (Agić and Vulić, 2019) from JW.org with over 343 languages, including over 100 African languages. Apart from the training dataset, evaluation datasets are needed to test the performance of multilingual MT models. The FLORES-101 (Goyal et al., 2021) evaluation set, sourced from Wikipedia and manually translated, covers the largest number of languages, including 20 African languages. Finally, while other evaluation datasets for translating into or from African languages have been developed (Siminyu et al., 2021; Emezue and Dosou, 2020; Azunre et al., 2021b; Nyoni and Bassett, 2021; Gezmu et al., 2021; Ali et al., 2021), unfortunately there are only a few African languages with evaluation datasets in the news domain (Adelani et al., 2021a; Mabuya et al., 2021; Ezeani et al., 2020) but ours covers 11 African languages (§4).

Low-resource MT. Interest in low-resource MT has been increasing both within the MT research community (Haddow et al., 2021), as well as in native speaker communities (∀ et al., 2020; Azunre et al., 2021a; Mager et al., 2021). On the modeling side, many techniques have been developed: unsupervised MT (Lample et al., 2018) leverages monolingual data, single multilingual models capable of translating between many languages (Firat et al., 2016; Johnson et al., 2017; Aharoni et al., 2021)
Table 1: Languages and Data Details for MAFAND-MT Corpus. Language, family (NC: Niger-Congo), number of speakers, news source, news (NEWS), and religious domain (REL) data split. The languages highlighted in gray did not previously have news-domain data before MAFAND-MT.

<table>
<thead>
<tr>
<th>Target Language</th>
<th>Family</th>
<th>African Region</th>
<th>No. of Speakers</th>
<th>Source Language</th>
<th>Source</th>
<th>NEWS</th>
<th>Split Sizes</th>
<th>REL</th>
<th>Source Total Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bambara (bam)</td>
<td>NC/Bantu</td>
<td>West</td>
<td>14M</td>
<td>French</td>
<td>Malawi.net</td>
<td>3302/ 14/144/ 1600</td>
<td>Bible</td>
<td>28K</td>
<td></td>
</tr>
<tr>
<td>Gomála’ (gbl)</td>
<td>NC/Grassfields</td>
<td>Central</td>
<td>1M</td>
<td>French</td>
<td>Cameroon Web</td>
<td>2233/ 113/130/ 1430</td>
<td>Bible</td>
<td>18K</td>
<td></td>
</tr>
<tr>
<td>Éwè (ewo)</td>
<td>NC/Kwa</td>
<td>West</td>
<td>7M</td>
<td>French</td>
<td>Renais Web TV</td>
<td>2020/ 14/144/ 1583</td>
<td>JW300</td>
<td>618K</td>
<td></td>
</tr>
<tr>
<td>Fon (fon)</td>
<td>NC/Volta-Niger</td>
<td>West</td>
<td>2M</td>
<td>French</td>
<td>ORTB, Nation, Hérit, Matin Libre, LB</td>
<td>2637/ 12/27/ 1579</td>
<td>JW300</td>
<td>32K</td>
<td></td>
</tr>
<tr>
<td>Hausa (hau)</td>
<td>Afro-Asiatic</td>
<td>East &amp; Central</td>
<td>63M</td>
<td>English</td>
<td>WMT2021: Khamenei.v1</td>
<td>3093/ 15/150/ 1500</td>
<td>JW300</td>
<td>258K</td>
<td></td>
</tr>
<tr>
<td>Igbo (ibo)</td>
<td>NC/Volta-Niger</td>
<td>West</td>
<td>27M</td>
<td>English</td>
<td>(Eseni et al., 2020)</td>
<td>6996/ 15/150/ 1500</td>
<td>JW300</td>
<td>415K</td>
<td></td>
</tr>
<tr>
<td>Lingala (ling)</td>
<td>NC/Bantu</td>
<td>East</td>
<td>7M</td>
<td>English</td>
<td>Bible</td>
<td>2026/ 14/144/ 1583</td>
<td>JW300</td>
<td>31K</td>
<td></td>
</tr>
<tr>
<td>Lolo (luo)</td>
<td>Nilo-Saharan</td>
<td>East</td>
<td>4M</td>
<td>English</td>
<td>Lukwe, Standard Media, Lufya, Lufya, Setswana</td>
<td>2426/ 15/150/ 1500</td>
<td>Bible</td>
<td>216K</td>
<td></td>
</tr>
<tr>
<td>Moun (mou)</td>
<td>NC/Gus</td>
<td>West</td>
<td>15M</td>
<td>French</td>
<td>Burkina24, Lefaso</td>
<td>2237/ 14/175/ 1774</td>
<td>JW300</td>
<td>318K</td>
<td></td>
</tr>
<tr>
<td>Naija (nten)</td>
<td>Nilo-Saharan-English-Creole</td>
<td>West</td>
<td>75M</td>
<td>English</td>
<td>Lomwe, Standard Media, Lufya, Setswana</td>
<td>2756/ 15/150/ 1500</td>
<td>JW300</td>
<td>214K</td>
<td></td>
</tr>
<tr>
<td>Swahili (swa)</td>
<td>NC/Volta-Niger</td>
<td>East &amp; Central</td>
<td>98M</td>
<td>English</td>
<td>Daily Trust Nigeria</td>
<td>4790/ 14/144/ 1584</td>
<td>JW300</td>
<td>23K</td>
<td></td>
</tr>
<tr>
<td>Setswana (tsn)</td>
<td>NC/Bantu</td>
<td>South</td>
<td>14M</td>
<td>English</td>
<td>Global Voices, OPUS</td>
<td>3670/ 17/191/ 1835</td>
<td>JW300</td>
<td>372K</td>
<td></td>
</tr>
<tr>
<td>akanTwi (twi)</td>
<td>NC/Kwa</td>
<td>West</td>
<td>9M</td>
<td>English</td>
<td>StarrFM, Citi News</td>
<td>3337/ 12/84/ 1500</td>
<td>JW300</td>
<td>601K</td>
<td></td>
</tr>
<tr>
<td>Wolof (wo)</td>
<td>NC/Senegalbou</td>
<td>West</td>
<td>5M</td>
<td>French</td>
<td>SeneWeb, Jotna, Yorun Post, Socialmedia</td>
<td>3860/ 15/150/ 1500</td>
<td>Bible</td>
<td>22K</td>
<td></td>
</tr>
<tr>
<td>yoruba (yor)</td>
<td>NC/Volta-Niger</td>
<td>West</td>
<td>42M</td>
<td>English</td>
<td>(Adelani et al., 2021a)</td>
<td>6644/ 15/144/ 1558</td>
<td>JW300</td>
<td>469K</td>
<td></td>
</tr>
<tr>
<td>isiZulu (zul)</td>
<td>NC/Bantu</td>
<td>South</td>
<td>27M</td>
<td>English</td>
<td>(Mabuya et al., 2021)</td>
<td>3500/ 12/39/ 998</td>
<td>JW300</td>
<td>667K</td>
<td></td>
</tr>
</tbody>
</table>

2Spoken by an estimated 1.1M people in Cameroon

2019; Fan et al., 2021a), multilingual unsupervised models leverage a related language (with parallel data) to assist translating the low-resource language that might not even have any monolingual data (Ko et al., 2021). Unfortunately, unsupervised MT typically performs poorly on low-resource languages (Marchisio et al., 2020).

Transfer learning from high-resource languages has achieved more promising results: Transfer from multilingual pre-trained language models (PLM), like mBART50 (Tang et al., 2020) and MT5 (Xue et al., 2021b), and large-scale multilingual MT often outperforms bilingual MT (Tran et al., 2021; Yang et al., 2021). For low-resource languages this strategy outperforms the baseline (Transformer) models (Birch et al., 2021; Adelani et al., 2021a; Lee et al., 2022). The performance can be further improved by large scale pre-training (Reid et al., 2021; Emezue and Dossou, 2021).

### 3 Focus Languages and Their Data

#### Focus Languages.

We focus on 16 African languages with varying quantities of available data (Joshi et al., 2020), including moderately low-resource languages such as Swahili and Hausa, and very low-resource languages such as Gomála’ with the Bible being its largest available corpus. Table 1 provides an overview of the focus languages, including the language families, location and number of speakers, and the source and original language for our corpus. The languages are from four language families: Afro-Asiatic (e.g. Hausa), Nilo-Saharan (e.g. Luo), English Creole (e.g. Nigerian-Pidgin/Naija) and Niger-Congo. Most of the languages (13 out of 16) are from the Niger-Congo family, which is the largest language family in Africa. Six of the languages are predominantly spoken in Francophone countries of Africa, while the remainder are predominantly spoken in Anglophone countries of Africa. In contrast to previous work (∀ et al., 2020; Gowda et al., 2021), we do not focus exclusively on translation to/from English since this is not the primary language of the Francophone Africa community. All languages are spoken by at least one million speakers.

#### Language Characteristics.

All languages are written in Latin script, using letters of the basic Latin alphabet with a few omissions (e.g “c”, “q”, “x”, “z”) and additions (e.g. “ë”, “ş”, “ţ”) and digraphs like “gb”, “kp”, “gh”, and sometimes more than two-character letters). 13 of the languages are tonal, and about nine make use of diacritics. Many African languages are morphologically rich. For example, all Bantu languages are aglutinative. Fon, Mossi, and Yorùbá are highly isolating. All languages follow the Subject-Verb-Object sentence structure like English and French. Table C provides more details.

#### Existing Parallel Corpora.

We curate publicly available parallel data for our focus languages, which consists primarily of text in the religious domain. For most African languages, the largest available parallel corpora is JW300 (Agič and Vulić, 2019), sourced from jw.org, which publishes biblical texts as well as lifestyle and opinion columns. Varying quantities of data are available for 11 of the 16 focus languages. Éwè, Igbo, Swahili, Setswana, Twi, Yorùbá, and isiZulu have over 400K parallel sentences. Hausa and Mossi have slightly more than 200K parallel sentences, while Fon and Naïja have around 30K sentences. For the remaining
five languages that are not in the JW300 corpus, we make use of the Bible. We aligned the sentences automatically by the verses (around 31k in total). Ghomá’ only has the New Testament with 8k verses. Bambara and Wolof are missing some verses and books, leading to a total size of 28K and 22K. Table 1 summarizes this information about the religious (REL) corpora.

4 MAFAND-MT African News Corpus

4.1 Data Collection Process

We introduce our newly translated news corpus; MAFAND-MT — Masakhane Anglo & Franco Africa News Dataset for Machine Translation. Table 1 gives the news source and data splits for 11 African languages which includes six languages (bam, bbj, ewe, fon, mos, wol) spoken predominantly in Francophone Africa and five languages (lug, luo, pcm, tsn, twi) spoken predominantly in Anglophone Africa. The MAFAND-MT corpus was created in three steps:

1. Crawling and preprocessing of news websites from local newspapers that are publishing in English and French. Raw texts from the web were segmented into sentences. Most languages were crawled from one or two sites, except for Wolof and Fon that were crawled from four and seven news websites respectively due to local French language newspapers having very few articles. We also ensured that the articles came from a variety of topics e.g. politics, sports, culture, technology, society, religion, and education. This was carried out by native speakers of the target language with source language proficiency.

2. Translation of 5k–8k sentences by professional translators. The translation process took one to four months depending on the availability of the translators.

3. Quality control was provided by native speakers, who discussed and, if possible, fixed problematic translations and ran automatic checks to detect misspellings, duplicated sentences, and alignment problems.

Following the recommendations of ∃ et al. (2020), we design the process to be participatory: Everyone involved in the corpus creation is a native speaker of the respective target languages and has societal knowledge about the communities that speak those languages. This is particularly important for curation and quality control to ensure that the resulting material is appropriate and relevant for stakeholders of the final MT models (∃ et al., 2020; Kreutzer et al., 2021). Furthermore, everyone received appropriate remuneration. To enable cross-disciplinary knowledge transfer between participants in the individual steps, every language was assigned a coordinator. The coordinator conducted the initial curation in the first step, and communicated with translators and quality checkers throughout the following steps.

Other Available Parallel Corpora. We found five African languages with available parallel texts in the news domain: Hausa5, Igbo (Ezeani et al., 2020), Swahili6, Yorùbá (Adelani et al., 2021a), and isiZulu (Mabuya et al., 2021). Table 1 provides news source, the TRAIN, DEV and TEST splits.

Appendix B provides details on the pre-processing of the available news corpora.

4.2 Monolingual News Corpus

To adapt available multilingual pre-trained models via continued pre-training to African languages, we curated texts from the 17 highest-resourced African languages and three non-native African languages that are widely spoken on the continent (Arabic, English, and French). The selection of African languages is based on their coverage in mC4 (Xue et al., 2021b), AfriBERTa corpora (Ogueji et al., 2021), and other publicly available news websites like VOA and BBC. We limited the size of the corpus extracted from mC4 to the first 30 million sentences (roughly 1GB of data) for Afrikaans, Amharic, Arabic, English, French, and Swahili. In total, we collected about 12.3 GB of data. Appendix C provides more details about the pre-training corpus.

5 Models and Methods

5.1 Baseline Models

We experiment with pre-trained multilingual models and our own bilingual MT baselines. We focus
We describe two methods for adding new languages under study are included. To cover 100 languages, less than half of the languages are included. As shown in Table 2, even for MT5 and M2M-100, which cover 418M parameters, both for computational feasibility and comparability across various different models.

**Transformer Baseline.** We train Transformer (Vaswani et al., 2017) sequence-to-sequence models from scratch for each language pair using JoeyNMT (Kreutzer et al., 2019). We tokenize the bitext using a joint SentencePiece unigram model (Kudo, 2018), with a character coverage of 1.0 and a maximum sentence length of 4096 tokens and create a vocabulary of 10K subwords. Models are trained on the concatenation of REL and NEWS corpora for each language.

**Pre-trained Models.** We consider three language models, MT5 (Xue et al., 2021b), ByT5 (a token-free T5) (Xue et al., 2021a), mBART50 (Tang et al., 2020), and the multilingual translation model M2M-100 (Fan et al., 2021b) for our experiments. We use MT5-base and ByT5-base, and M2M-100 with 418M parameters. Table 2 gives the pre-trained model size, number of African languages covered, and the focus languages supported.

<table>
<thead>
<tr>
<th>Model (PM)</th>
<th>Pre-trained Model (PM)</th>
<th>PM Size</th>
<th># African Lang.</th>
<th>Focus languages covered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afri[*T5]</td>
<td>580M</td>
<td>13</td>
<td>hau, ibo, swa, yor, zul</td>
<td></td>
</tr>
<tr>
<td>mBART50</td>
<td>610M</td>
<td>2</td>
<td>swa</td>
<td></td>
</tr>
<tr>
<td>AfriMBART</td>
<td>610M</td>
<td>17</td>
<td>hau, ibo, pcm, swa, yor, zul</td>
<td></td>
</tr>
<tr>
<td>M2M-100</td>
<td>418M</td>
<td>17</td>
<td>hau, ibo, lug, swa, tsn, wol, yor, zul</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Language coverage and size for pre-trained models. Afri[*T5] refers to AfriMT5/ByT5.

on pre-trained models that are approximately 500M parameters, both for computational feasibility and comparability across various different models.

**5.2 Transfer Learning Across Languages**

We describe two methods for adding new languages to existing models: continual pre-training and many-to-many multilingual translation.

**Continual Pre-training.** The effectiveness of PLMs is limited on extremely low-resource languages because they rarely, if ever, occur in the pre-training corpus (Wang et al., 2020; Liu et al., 2021). As shown in Table 2, even for MT5 and M2M-100, which cover 100 languages, less than half of the African languages under study are included. To adapt the existing PLMs to our languages corpora and domains, we apply continual pre-training (Gururangan et al., 2020; Liu et al., 2021) using our collected monolingual corpus. Specifically, before fine-tuning on the parallel MT data, models are pre-trained with their original training objective and vocabulary on the monolingual corpus. Pre-training parameters can be found in the appendix. We refer to the models adapted to African languages as AfriMT5, AfriByT5, and AfriMBART.

**Many-to-Many Translation.** We fine-tuned M2M-100 for African multilingual translation to create English- and French-centric models. For the English-centric model, the M2M-100 model was fine-tuned on the news data for en–{hau, ibo, lug, luo, pcm, swa, tsn, twi, yor, zul} while the French-centric model is trained on fr–{bam, bbj, ewe, fon, mos, wol}. Languages not included in the pre-trained M2M-100 model were assigned the language code of a language included in M2M-100 but excluded from our study.

**5.3 Transfer Learning Across Domains**

As there is very limited MT data on the news domain, we compare different methods that combine the large data from the religious domain (REL) and the small data from the NEWS domain (NEWS) to fine-tune M2M-100:

1. REL+NEWS: Fine-tuning on the aggregation of REL and NEWS.
2. REL–>NEWS: Training on REL, followed by fine-tuning on NEWS.
3. REL+NEWS–>NEWS: REL+NEWS, followed by additional fine-tuning on NEWS.

Each fine-tuning stage lasts for three epochs. We evaluate translation quality with BLEU (Papineni et al., 2002) using SacreBLEU (Post, 2018) and ChrF (Popović, 2015).

**6 Results and Discussion**

We successfully adapt several multilingual pre-trained models to previously unseen African languages and quantify the effectiveness of small in-domain translation datasets. We discuss the effects of domain shift and analyze mitigation strategies.

### 6.1 Adaptation to the Focus Languages

We demonstrate that fine-tuning with a few thousand high-quality bitext is effective for adding new languages to pre-trained models. Further, continuing to fine-train to specialize models to African languages further improves performance.

---

7https://github.com/google/sentencepiece

---

Chang the vocabulary (Gururangan et al., 2020) to fit the languages, or adding MT-focused training objectives for word alignment (Liu et al., 2021) can potentially improve the performance further, which we leave for future work.

9“intl” tokenizer, all data comes untokenized.
Zero-Shot Translation. Table 3 and Table 4 gives the result of zero-shot evaluation on NEWS. We evaluate only on the M2M-100 dataset because it has been pre-trained on parallel texts with a few of our focus languages. We observe very poor performance (< 5 BLEU) on the languages except for Zul ( > 13 BLEU) and swa ( > 20 BLEU) in both translation directions. For swa, its likely that the performance is reasonable because M2M-100 has seen more bitext during pre-training (2.4M sentences in CCAligned (El-Kishky et al., 2020)). Other African languages except for Afrikaans have less than 600K sentences in CCAligned, and are also of a lower quality (Kreutzer et al., 2021) which affect overall zero-shot performance.

Performance after Fine-tuning. We found impressive performance after fine-tuning PLMs and M2M-100 on few thousand sentences (mostly 2K–7K sentences, except for swa with 30K sentences), including languages not seen during pre-training. For en/fr-xx, MT5 has a poor transfer performance with average BLEU of 7.2, despite being pre-trained on 101 languages. ByT5 outperforms MT5 by over 3 BLEU on average, even though their performances were reported to be similar in previous work (Xue et al., 2021a). This indicates that ByT5 might be preferable over MT5 when translating low-resource languages. Surprisingly, mBART5 that was only pre-trained on 50 languages and 2 African languages outperformed MT5 and ByT5 which are pre-trained on 101 languages. Overall, we found M2M-100 to be the best model, most likely because it was pre-trained on a translation task. In general, BLEU scores are relatively low (< 15 BLEU for 9 out of 16 languages for en/fr-xx and 7 in xx-en/fr) even when fine-tuning M2M-100 on in-domain data, which suggests that developing more effective methods for fine-tuning might be a promising future direction. The languages with the best quality according to BLEU on the target side are pcm, swa and tsn, and pcm, zul, and swa on the source side.

BLEU scores are higher when translating from an African language, which is expected due to the more frequent exposure to English and French on the target side during pre-training, and BLEU being penalized more for morphologically rich languages like bjj, lug, swa, tsn, and zul, the ChrF metric works better for them. For example, fine-tuning M2M-100 on NEWS and evaluating on zul has a BLEU of 21.0 in enfr-xx, and BLEU of 37.8 in the xx-en/fr showing a large gap in performance in both directions. However, with the ChrF, we find a smaller performance gap (51.2 in enfr-xx and 55.5 in the xx-en/fr).

Continual Pre-training. We observe an improvement in BLEU when we utilize AfriMT5 and AfriByT5, for languages included in our continual pre-training corpus (Appendix C). Other languages also benefit despite not being seen during continual pre-training, possibly due to language similarity. For example, AfriByT5 on fr-bam improved by 1.9 BLEU over ByT5 and AfriMT5 on en-tns improved by 3.6 BLEU over MT5. On average, AfriMT5 im-
proved over MT5 by 1.3 BLEU in en/fr-xx and 2.4 BLEU in the xx-en/fr. The improvement for AfriByT5 was much smaller: 0.6 and 0.9 BLEU in en/fr-xx and xx-en/fr translation directions. For AfriMBART, we did not see any improvement on average, only the performance of hau (1.5 BLEU) and ibo (0.7 BLEU) improved in en/fr-xx direction. However, in the xx-en/fr direction, fon, tsn, twi, and zul improved by 2.7–6.0 BLEU.

Many-to-Many Multilingual MT. Training on the combined news corpus from all languages that use French or English separately does not appear to help much. We see slight improvements for most languages only in the xx-en/fr direction.

6.2 Adaptation to the News Domain

To improve over the baseline performance on NEWS, we train bilingual Transformer models (as a baseline) and M2M-100 on a combination of REL and NEWS. We chose M2M-100 because it was the best performing model. Table 5 gives the BLEU on three settings: REL+NEWS, REL→NEWS, and REL+NEWS→NEWS. In general, the improvement depends on the size of REL corpus. For languages trained on the Bible such as bbj, bam, lug, lu, and wol, the improvement is minimal. For M2M-100, the REL+NEWS performance does not improve over NEWS despite the larger quantity of training data. This demonstrates that increasing the size in the target domain is the most helpful strategy (see Figure 2). Similarly, combining REL+NEWS is not very helpful for xx-en/fr. An alternative approach is REL→NEWS, which allows the model to develop a good understanding of the desired language before adapting to the news domain. We observe an increase on 1.1 BLEU over REL+NEWS in the en/fr-xx direction. However, the best strategy is REL+NEWS→NEWS, especially for xx-en/fr where it yields an improvement over NEWS and REL+NEWS by 2.0 and 1.5 BLEU, respectively.

6.3 Analysis of Domain Shift

Is a small in-domain set essential for fine-tuning? If we train models only on previously available religious data, they are not capable of translating news well due to the strong domain bias. This is illustrated in Figure 1: All models perform much worse on NEWS than on the REL do-

Table 4: Results adding African Languages to Pre-Trained Models, xx-en/fr. We calculate BLEU and CHRF on the news domain when training on only NEWS data from MAFAND-MT.

<table>
<thead>
<tr>
<th>Model</th>
<th>bam</th>
<th>bbj</th>
<th>ewe</th>
<th>fon</th>
<th>mos</th>
<th>wol</th>
<th>xx-en</th>
<th>xx-fr</th>
<th>AVG</th>
<th>MED</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M2M-100 0-shot</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MT5</td>
<td>2.5</td>
<td>0.9</td>
<td>1.1</td>
<td>2.4</td>
<td>0.7</td>
<td>1.3</td>
<td>0.8</td>
<td>0.6</td>
<td>2.2</td>
<td>6.4</td>
</tr>
<tr>
<td>AfriMT5</td>
<td>6.4</td>
<td>2.0</td>
<td>2.1</td>
<td>4.2</td>
<td>1.2</td>
<td>2.9</td>
<td>10.4</td>
<td>19.5</td>
<td>15.5</td>
<td>9.7</td>
</tr>
<tr>
<td>ByT5</td>
<td>10.0</td>
<td>2.7</td>
<td>4.1</td>
<td>4.9</td>
<td>1.5</td>
<td>7.2</td>
<td>12.9</td>
<td>21.0</td>
<td>19.8</td>
<td>12.1</td>
</tr>
<tr>
<td>AfriByT5</td>
<td>13.8</td>
<td>4.4</td>
<td>4.5</td>
<td>5.8</td>
<td>2.2</td>
<td>9.0</td>
<td>15.3</td>
<td>20.7</td>
<td>21.3</td>
<td>12.5</td>
</tr>
<tr>
<td>mBART50</td>
<td>6.8</td>
<td>0.3</td>
<td>1.7</td>
<td>0.8</td>
<td>0.6</td>
<td>6.3</td>
<td>11.5</td>
<td>13.2</td>
<td>9.1</td>
<td>4.4</td>
</tr>
<tr>
<td>AfriMBART</td>
<td>8.1</td>
<td>2.3</td>
<td>3.0</td>
<td>4.5</td>
<td>1.7</td>
<td>3.2</td>
<td>10.2</td>
<td>15.5</td>
<td>13.1</td>
<td>8.0</td>
</tr>
</tbody>
</table>

| BLEU        |     |     |     |     |     |     |     |     |     |     |
| M2M-100     |     |     |     |     |     |     |     |     |     |     |
| M2M-100-ENFR|     |     |     |     |     |     |     |     |     |     |
| MT5         | 19.4| 15.1| 17.0| 17.9| 10.9| 16.2| 26.3  | 41.5  | 36.3 | 26.1 |
| AfriMT5     | 27.7| 19.6| 21.1| 21.4| 13.2| 21.6| 32.5  | 44.9  | 40.2 | 32.2 |
| ByT5        | 31.2| 21.8| 24.8| 20.5| 15.4| 26.2| 31.3  | 46.4  | 45.4 | 34.1 |
| AfriByT5    | 34.8| 25.5| 24.9| 22.0| 16.2| 29.3| 33.9  | 46.4  | 47.1 | 35.0 |
| mBART50     | 26.0| 17.1| 20.9| 20.2| 17.1| 26.6| 32.0  | 37.9  | 39.0 | 31.0 |
| AfriMBART   | 31.4| 22.9| 27.2| 26.3| 17.0| 25.0| 34.3  | 42.0  | 40.4 | 29.8 |
| M2M-100     | 45.9| 26.5| 30.9| 27.5| 17.7| 33.8| 48.7  | 46.1  | 46.4 | 36.7 |
| M2M-100-ENFR| 45.6| 26.9| 32.2| 28.7| 17.0| 34.3| 35.1  | 46.6  | 46.0 | 37.6 |

Figure 1: Domain shift of M2M-100 Transformer models trained on en/fr-xx (top) or xx-en/fr (bottom) REL domain and tested on the NEWS vs. REL domains.
Table 5: Results adapting to Domain Shift, en/fr-xx. We calculate BLEU and ChrF on the news domain when training on different combinations of REL and NEWS.

<table>
<thead>
<tr>
<th>Model</th>
<th>ham</th>
<th>bblj</th>
<th>ewe</th>
<th>fon</th>
<th>mos</th>
<th>wol</th>
<th>hau</th>
<th>ibo</th>
<th>lug</th>
<th>luo</th>
<th>pcm</th>
<th>swa</th>
<th>tsn</th>
<th>twi</th>
<th>yor</th>
<th>zul</th>
<th>AVG</th>
<th>MED</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>REL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transformer REL+NEWS</td>
<td>4.0</td>
<td>4.8</td>
<td>4.3</td>
<td>4.2</td>
<td>4.1</td>
<td>4.0</td>
<td>3.9</td>
<td>3.8</td>
<td>3.7</td>
<td>3.6</td>
<td>3.5</td>
<td>3.4</td>
<td>3.3</td>
<td>3.2</td>
<td>3.1</td>
<td>3.0</td>
<td>2.9</td>
<td>2.8</td>
</tr>
<tr>
<td>REL+NEWS→ENG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>REL+NEWS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transformer REL+NEWS</td>
<td>4.0</td>
<td>4.8</td>
<td>4.3</td>
<td>4.2</td>
<td>4.1</td>
<td>4.0</td>
<td>3.9</td>
<td>3.8</td>
<td>3.7</td>
<td>3.6</td>
<td>3.5</td>
<td>3.4</td>
<td>3.3</td>
<td>3.2</td>
<td>3.1</td>
<td>3.0</td>
<td>2.9</td>
<td>2.8</td>
</tr>
<tr>
<td>REL+NEWS→ENG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Results adapting to Domain Shift, xx-en/fr. We calculate BLEU and ChrF on the news domain when training on different combinations of REL and NEWS.

<table>
<thead>
<tr>
<th>Model</th>
<th>ham</th>
<th>bblj</th>
<th>ewe</th>
<th>fon</th>
<th>mos</th>
<th>wol</th>
<th>hau</th>
<th>ibo</th>
<th>lug</th>
<th>luo</th>
<th>pcm</th>
<th>swa</th>
<th>tsn</th>
<th>twi</th>
<th>yor</th>
<th>zul</th>
<th>AVG</th>
<th>MED</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>REL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transformer REL+NEWS</td>
<td>4.0</td>
<td>4.8</td>
<td>4.3</td>
<td>4.2</td>
<td>4.1</td>
<td>4.0</td>
<td>3.9</td>
<td>3.8</td>
<td>3.7</td>
<td>3.6</td>
<td>3.5</td>
<td>3.4</td>
<td>3.3</td>
<td>3.2</td>
<td>3.1</td>
<td>3.0</td>
<td>2.9</td>
<td>2.8</td>
</tr>
<tr>
<td>REL+NEWS→ENG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>REL+NEWS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transformer REL+NEWS</td>
<td>4.0</td>
<td>4.8</td>
<td>4.3</td>
<td>4.2</td>
<td>4.1</td>
<td>4.0</td>
<td>3.9</td>
<td>3.8</td>
<td>3.7</td>
<td>3.6</td>
<td>3.5</td>
<td>3.4</td>
<td>3.3</td>
<td>3.2</td>
<td>3.1</td>
<td>3.0</td>
<td>2.9</td>
<td>2.8</td>
</tr>
<tr>
<td>REL+NEWS→ENG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7: Example translations for M2M-100 fine-tuned on REL or REL+NEWS→NEWS (R+N→N). Terms in red are typical for biblical texts, while the terms in blue are more neutral expressions.

<table>
<thead>
<tr>
<th>lang</th>
<th>src</th>
<th>tr</th>
</tr>
</thead>
<tbody>
<tr>
<td>bam</td>
<td>Ini ka fu ye ko nswen.</td>
<td>Ini ka fu ye ko nswen.</td>
</tr>
<tr>
<td>fr</td>
<td>Et leur dire que la transition se repose sur trois piliers.</td>
<td>Et leur dire que la transition se repose sur trois piliers.</td>
</tr>
<tr>
<td>eng-bz</td>
<td>You don’t have to know everything in advance.</td>
<td>You don’t have to know everything in advance.</td>
</tr>
<tr>
<td>eng-sw</td>
<td>And Murshash Shikib wrote a complete book first in the world.</td>
<td>And Murshash Shikib wrote a complete book first in the world.</td>
</tr>
</tbody>
</table>

main. When the quantity of religious training data is small, the loss in translation performance on the news test set is largest, c.f. bbj (8k of REL data) with a drop of -95.5% BLEU or bam (-93.5%, 28k) and luo (-93.5%, 31k). This indicates that when the REL training data is sparse, it is insufficient to teach the M2M-100 model a more general understanding required for translating NEWS. However, when the religious training data is larger, this loss is reduced, c.f. when translating to zul (667k, -67%), swa (-69.3%, 872k), and tsn (-71%, 870k). While this is the general trend, pcm, whose religious training data is small (23k), has the lowest drop in performance (-59.3%), which may be due to the strong similarity to its source language.

How many sentences in the target domain are required? Figure 2 shows how for three selected language pairs with a large (fr-bam), medium (eng-bz) and relatively small (eng-sw) domain gap, the quality of target domain translations improves as we increase the size of the target domain corpus. For all three pairs, fine-tuning M2M-
100 or ByT5 on 2.5k sentence pairs of in-domain data (NEWS) is sufficient to outperform the bilingual Transformer baselines that were additionally trained on larger amounts of out-of-domain data (REL). Surprisingly, this procedure not only works for languages included during pre-training (swa), but also for previously unseen languages (ibo, bam). M2M-100 tends to adapt to the new data more quickly than ByT5, but in all cases, models continue to learn with additional in-domain data. This shows how much more effectively a small number of in-domain translations can be used when they serve for fine-tuning multilingual pre-trained models rather than training bilingual MT models from scratch.

### Examples of Domain Bias

To illustrate the challenge of overcoming domain bias, we show examples translating from bam and lug in Table 7. The M2M-100 model fine-tuned only on REL succeeds in roughly capturing the meaning of the sources, but using biblical terms, such as “scroll” instead of “novel”. Adding our news corpus to fine-tuning resolves these issues (e.g. “book”).

### How general is our news corpus?

Table 8 shows the zero-shot evaluation of M2M-100 fine-tuned on our small NEWS corpora on other domains: religious (REL) and Wikipedia (FLORES). We evaluated the Wikipedia domain on the FLORES devtest and the REL domain on either JW300 or Bible (lug, luo, wol). As a baseline, we evaluated the zero-shot performance of M2M-100 (not fine-tuned, ✗) on FLORES\(^{10}\) using spBLEU (i.e. sentencepiece BLEU (Goyal et al., 2021)). We noticed very poor performance except for Swahili — as discussed in §6.1. After fine-tuning on our new data (✓), transfer is largely improved across the bench (up to +17 BLEU for en–ibo). The same trend holds for the religious domain. This shows that even though our data comes from the news domain, it helped the model generalize to other domains. Hence, expanding African news corpora and developing better MT models for news pays off even for other domains of interest.

### 7 Conclusion

We have created MAFAND-MT, a corpus of 16 African languages to study translation systems for low-resource languages in the news domain. We investigate how to most effectively adapt large-scale pre-trained models to incorporate new languages and new domains. Our findings suggest that as little as 2k sentences are sufficient for fine-tuning, with an improved performance, paving the way for others to create new translation systems without relying on large collections of web-sourced text. This has strong implications for languages that are spoken by millions but lack presence on the web.

In the future, we hope to expand our coverage to additional under-resourced languages, and to develop even more effective fine-tuning objectives. Currently, we are extending our corpus to Chichewa, Kinyarwanda, Shona, and isiXhosa, including an expansion of the Hausa corpus, they will be released under MAFAND-MT dataset name.

---

\(^{10}\)except for Luo which is not supported
8 Acknowledgment

This work was carried out with support from Lacuna Fund, an initiative co-founded by The Rockefeller Foundation, Google.org, and Canada’s International Development Research Centre. David Adelani acknowledges the EU-funded Horizon 2020 projects: COMPRISEx (http://www.compriseh2020.eu/) under grant agreement No. 3081705 and ROXANNE under grant number 833635. We thank Chester Chester Palen-Michel and Constantine Lignos for providing the VOA corpus for this research, and Google for providing GCP credits to run some of the experiments. Finally, we thank Davor Orlić and Knowledge4All for their administrative support throughout the project.

References


Alexandra Birch, Barry Haddow, Antonio Valerio Miceli Barone, Jindrich Hecl, Jonas Waldendorf, Felipe Sánchez Martínez, Mikel Forcada, Víctor Sánchez Cartagena, Juan Antonio Pérez-Ortiz, Miquel Esplà-Gomis, Wilker Aziz, Lina Murady,


Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks. In


Roowetheir Mabuya, Jade Abbott, and Vukosi Marivate. 2021. Umsuka english - isizulu parallel corpus. Thank you to Facebook Research for funding the creation of this dataset.


Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luís Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando


**A Language Characteristics**

Table 9 provides the details about the language characteristics.

**B Available Parallel Corpora**

We found Five African languages with publicly available parallel texts in the news domain: Hausa, Igbo, Swahili, Yorubá, and isiZulu. Table 1 provides news source, the Train, Dev and Test splits.

**Hausa** The Hausa Khamenei corpus contains 5,898 sentences, we split them into Train (3,098), Dev (1,300), and Test split (1,500).

**Igbo** The Igbo corpus (Ezeani et al., 2020) has 9,998 sentences, we extract 6,998 sentences for Train, and the remaining for Dev and Test splits.

**Swahili** The Global Voices corpus contains 30,782 sentences, which we use for the Train split. We additionally crawled newer (2019–2021) publications of Swahili articles from the Global Voices website, this gives a total of 3,626 sentences, they were aligned and manually verified by Swahili speakers. They are split into the Dev and Test splits.

**Yorùbá** The MENYO-20k (Adelani et al., 2021a) corpus contains sentences from different domains (TED talks, books, software localization, proverbs, and news), from which we select the news domain sentences for the Train, Dev and Test splits.

**isiZulu** The Umsuka corpus (Mabuya et al., 2021) contains 9,703 training sentences and 1,984 evaluation sentences. 4,739 training sentences were translated from English-isiZulu, and the remaining from isiZulu-English. We only keep the training sentences translated into isiZulu, and split them into 3,500 for Train and 1,239 sentences for Dev. From the existing evaluation set we select only the 998 English-isiZulu translations for Test. Umsuka provides two translations for each English sentence, but we use only the first.

**C Monolingual Corpus PLMs adaptation**

Table 10 provides the details about the Monolingual corpus used to adapt the pre-trained language models (PLMs), their size and source of corpora. The African languages pre-trained are: Afrikaans, Amharic, Hausa, Igbo, Malagasy, Chichewa, Oromo, Naija, Kinyarwanda, Kirundi.

---

11https://www.statmt.org/wmt21/translation-task.html
12https://sw.globalvoices.org/
Shona, Somali, Sesotho, Swahili, isiXhosa, Yorùbá, and isiZulu.

D Model Hyper-parameters and Reproducibility of Results

For the pre-trained models, we fine-tune the models using HuggingFace transformer tool (Wolf et al., 2020) with the default learning rate ($5e^-5$), batch size of 10, maximum source length & maximum target length of 200, beam size of 10, and number of epochs is 3 except for models trained on only NEWS which we set to 10. We make all the experiments were performed on a single GPU (Nvidia V100).

For fine-tuning pre-trained models, especially for mBART50 that only supports two African languages, the target language is required to be specified during decoding from among those that the model has seen during pre-training, we follow past works (Madaan et al., 2020; Cahyawijaya et al., 2021; Lee et al., 2022) in selecting another closely-related language that is represented in the pre-trained model. For convenience, we make use of Swahili (sw) as the target language when an African language is not represented since Swahili is represented in all the pre-trained models. The only exception is Nigerian-Pidgin, where we make use of French (fr) since it is closely related to English. When a language is represented in the pre-trained model like M2M-100 has seen Yorùbá (yo), we follow past works (Madaan et al., 2020; Cahyawijaya et al., 2021; Lee et al., 2022) in selecting another closely-related language that is represented in the pre-trained model. For convenience, we make use of Swahili (sw) as the target language when an African language is not represented since Swahili is represented in all the pre-trained models. The only exception is Nigerian-Pidgin, where we make use of French (fr) since it is closely related to English. When a language is represented in the pre-trained model like M2M-100 has seen Yorùbá (yo), we make use of the correct language code.

To train AfriMT5 and ByT5, we start with MT5 and ByT5. We pre-train with the learning rate $1e^-4$, 10,000 warm up steps and a batch size of 2048 for one epoch. For mBART50, we pre-train with learning rate $5e^-5$ for 50,000 steps.
We observe that spBLEU gives higher scores than AfriByT5, and AfriMBART. The following examples from the Fon-to-French translation of the test set illustrate the advantage of multilingual modeling and its limitations:

xx-en/fr

en/fr-xx

AfriM5

AfriByT5

AfriMBART

mBART50

AfriMT5

AfriByT5

AfriMBART

M2M-100

M2M-100-EN

M2M-100-FR

REL

REL+NEWS

REL+NEWS=NEWS

xx-en/fr

en/fr-xx

AfriMT5

AfriByT5

AfriMBART

M2M-100

M2M-100-EN

M2M-100-FR

Table 11: Model names on HuggingFace Model Hub. For bilingual models, supply the correct src or tgt language. English/French make use of a 2-letter language code i.e. en or fr, while all the African languages make use of 3-letter language codes e.g. yor.

Table 12 and Table 13 compares BLEU and sp-BLEU metric for the domain transfer experiments. We observe that spBLEU gives higher scores than BLEU especially in the direction of en/fr-xx, which shows that it may be better for evaluating African languages. Although, further analysis and human evaluation are still needed to show that spBLEU is generally better. On the other hand, in the xx-en/fr, there is no much difference in the scores between BLEU and spBLEU.

Table 12: BLEU on Wikipedia domain (FLORES), REL, and NEWS for M2M-100 before (✓) and after (✓) fine-tuning on NEWS.

E BLEU vs spBLEU

using Fairsseq (Ott et al., 2019) without modifying the mBART50 vocabulary. Table 11 has the names of all the models that are publicly available on HuggingFace Model Hub. In total, we have 357 models from 22 x 16 bilingual models, two English/French-centric models, and three adapted models to African languages (i.e. AfriM5, AfriByT5, and AfriMBART).

F Qualitative Analysis

The following examples from the Fon-to-French translations of the test set illustrate the advantage of multilingual modeling and its limitations:

- **Source** (fon): Louis Guy Alimanyidokpo kpóössà Ètchlekoun kpó ø, sìn òzàn mokpàn dyè ø, ye òwèr sè we twàwn dyè agbàza mè, cò ye kà tuun fi è àzon nè lèy gòsin è øà.

- **Reference** (fr): Les faits Louis Guy Alimagnidokpo et Issa Ètchlekoun se plaignent

13https://huggingface.co/masakhane

since some days of multiple douleurs, ignoring the origine réelle de leurs maux.

- **Bilingual Transformer** (REL+NEWS, fon→fr): on ne peut pas avoir une trentaine d’années ni un jeune homme ni un jeune homme d’âge pour un jeune homme qui soit 12 ans.

- **M2M-100** (REL+NEWS→NEWS, fon→fr): Sìn azàn yëwërywë dyè èkòpò wë nù è kàn Louis Guy Alimagnidokpo kpóössà Ètchlekùn kèpàn dè òdò dyø wë dyè wëvër gege wë, ye kà tuun nù è wë wëvër yeton dè øà.

The translation of the bilingual Transformer model
Despite the promising results, our work has the following limitations:

1. **Translation quality**: Even the best model scores low BLEU on some of the reported languages (bbj, mos, zul), in particular when translating into them.

2. **Evaluation**: Our evaluation is focused on BLEU. We report ChrF results as well, but without a deeper human evaluation, we cannot make claims about the absolute quality of the translations. Manual inspections of translations like the example discussed in Section F gave us the impression that translations are surprisingly fluent and make good use of language-specific expressions when translating into English or French, but that errors in grammar and logic can be easily overlooked. Automatic reference-based metrics like BLEU and ChrF might not be able to capture the semantic relatedness to the reference sufficiently, as well potentially being tricked by word matches in incoherent phrases.

3. **Language bias**: We have shown that even when not included in pre-training, and without large out-of-domain data, significant gains in translation quality can be achieved. However, language-specific biases, in terms of resourcefulness, morphology, standardization, inclusion in pre-trained models and available corpora, or relatedness to other languages, still affect the relative quality of translations, and require more efforts to be overcome.

4. **Domain limitations**: While we showed a rapid adaptation to the news domain and the auxiliary benefit of the religious domain, our study also revealed how automatically estimated translation quality drops when the test domain is narrow. Therefore, future work should aim to expand the study to multiple test domains and develop systematic methods.

### Table 13: \textit{spBLEU on Wikipedia domain} (FLORES), REL, and NEWS for M2M-100 before (✗) and after (✓) fine-tuning on NEWS.

<table>
<thead>
<tr>
<th>Evaluation Domain</th>
<th>Tuned on NEWS</th>
<th>haun</th>
<th>ibos</th>
<th>lugu</th>
<th>luos</th>
<th>swa</th>
<th>woli</th>
<th>yor</th>
<th>zuli</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.75 - 1.50 FLORES</td>
<td>✓</td>
<td>0.8</td>
<td>7.2</td>
<td>3.7</td>
<td>-</td>
<td>26.9</td>
<td>3.0</td>
<td>3.8</td>
<td>11.9</td>
</tr>
<tr>
<td>FLORES</td>
<td>16.9</td>
<td>12.0</td>
<td>7.7</td>
<td>11.8</td>
<td>25.8</td>
<td>7.5</td>
<td>9.3</td>
<td>19.2</td>
<td></td>
</tr>
<tr>
<td>REL</td>
<td>✓</td>
<td>6.4</td>
<td>3.7</td>
<td>0.5</td>
<td>-</td>
<td>15.4</td>
<td>0.4</td>
<td>0.9</td>
<td>8.5</td>
</tr>
<tr>
<td>NEWS</td>
<td>✓</td>
<td>3.8</td>
<td>0.0</td>
<td>1.7</td>
<td>2.5</td>
<td>13.9</td>
<td>1.7</td>
<td>5.7</td>
<td>12.5</td>
</tr>
<tr>
<td>NEWS</td>
<td>✓</td>
<td>2.6</td>
<td>9.1</td>
<td>7.2</td>
<td>2.7</td>
<td>27.8</td>
<td>1.0</td>
<td>3.9</td>
<td>15.7</td>
</tr>
</tbody>
</table>

In the opposite translation direction, \( \text{fr} \rightarrow \text{fon} \), M2M-100 ({REL + NEWS} → NEWS) still preserved some sense of logical reasoning and predicted the last part right ye kà tuun nû wà wù wàwè yèwà (they do know why they are suffering) dè ʒà (NOT). However, the model had some limitations: the names which are part of the translation are not spelled correctly. Some expressions are incomplete: For instance  \( \text{sin azàn} + \text{number} \) means since \text{xxx days} but yrywë is not a number, and do not have any meaning in this context.

### Language bias

- **Wrong Word**: The French reference made use of the word \( \text{wàwè} \) means \text{souffrir, souffrance} (suffering). The French reference made use of \( \text{se plaignent} \) (complaining) which makes less sense than \text{souffert} used in the M2M-100 prediction. M2M-100 also learned the style of the sentence: \( \text{çô ye kà tuun fi é azàn nû lëz gosin (but they do know the origin of their sufferings) é ʒà (NOT)} \) - this last part is crucial for the meaning of the entire sentence. Given the structural and morphological differences between Fon and French, we expected it to be more complicated to predict. However, this translation is structurally wrong even though any French native speaker would understand the conveyed message quickly and easily. In the M2M-100 translation, the word \text{malgrè} is at the wrong place, corrupting syntax and logic of the second clause. A perfect translation (in the idea to be expressed) would be: "Louis Guy Alimanyion et Issa Etchlekon ont depuis plusieurs jours souffert d’une maladie grave malgrè (don’t) ils ne connaissent pas les conséquences (causes/raisons) de cette maladie qu’ils ne connaissent pas."

- **Wrong Placement**: In the opposite translation direction, \( \text{fr} \rightarrow \text{fon} \), M2M-100 ({REL + NEWS} → NEWS) still preserved some sense of logical reasoning and predicted the last part right ye kà tuun nû wà wù wàwè yèwà (they}
for distilling knowledge from multiple narrow domains.

5. **Language coverage**: Africa has thousands of other languages that are not covered in our study but deserve the same attention. We hope that our work is encouraging enough to inspire native speakers of those languages not covered here to collect translations, run our code, and report their findings to the NLP research community, so that we can make joint progress in developing language technology for more people.

We believe that our translation models carry similar risks of causing harm by inaccurate and biased translations as the underlying large pre-trained models. M2M-100 is trained on large collections of texts crawled from the web, and the quality for most of the languages studied here is questionable (Kreutzer et al., 2021). Our fine-tuning successes show that some obvious biases can be overcome when the quality of the fine-tuning set is controlled (see the examples in Section 6.3), but we cannot guarantee that biases prevailing in the pre-training corpus or more subtle biases will not occur with other inputs. Together with a careful human evaluation, this should be the main concern for future work on the produced models. The methodology of rapid fine-tuning might also be misused to tune the models towards harmful content or purposes that harm the speakers of the languages presented here.